Lesion Analysis from Ultra-Low to Ultra-High Field MRI in Multiple Sclerosis

Corinne Alison Donnay

Linacre College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Michaelmas 2025

Abstract

Ultra-high field (UHF) and ultra-low field (ULF) magnetic resonance imaging (MRI) offer complementary approaches for studying and monitoring white matter lesions (WMLs) in multiple sclerosis (MS). UHF MRI provides enhanced sensitivity to chronic lesion pathology, while ULF MRI enables portable and accessible imaging. Existing image analysis tools, developed primarily for conventional MRI, often fail at these field strengths. This thesis develops and applies dedicated image analysis tools for WMLs at UHF (7T) and ULF (64mT), addressing field-specific research questions.

In the first part of this thesis, 7T tools were developed to quantify longitudinal changes in chronic lesions. Three analytical tools were developed: a 3T–7T transfer learning model for whole-brain and WML segmentation; an automated pipeline for tracking lesion quantitative T1 and volume changes; and a web-based application for standardized radiological lesion assessment. Application to a large longitudinal natural history cohort and two clinical trials revealed new insights into the long-term evolution of chronic lesions with and without paramagnetic rims, while establishing important methodological considerations for implementing 7T-derived biomarkers in treatment monitoring.

The second part addresses ULF MRI tool development to determine whether ULF imaging can capture clinically relevant lesion metrics compared to a 3T clinical standard. This work involved adapting transfer learning models for ULF WML segmentation, implementing a longitudinal subtraction pipeline to support radiological detection of evolving lesions, and evaluating a Fourier-based super-resolution approach combining orthogonal acquisitions to improve image quality. Analysis of paired 3T–64mT data demonstrated that 3T-derived lesion metrics and their respective associations with clinical disability could be captured at ULF. These findings contribute to the characterization of ULF's sensitivity and detection limits for WML, thereby informing its potential clinical utility.

Together, this research demonstrates that leveraging UHF and ULF can deepen our understanding of lesion pathophysiology and extend the reach of WML imaging in MS.

Keywords: Multiple sclerosis, MRI, ultra-low field, ultra-high field, white matter lesions, paramagnetic rim lesions

Lesion Analysis from Ultra-Low to Ultra-High Field MRI in Multiple Sclerosis



Corinne Alison Donnay

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2025

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Ludovica Griffanti, Dr. Daniel S. Reich, and Professor Heidi Johansen-Berg, whose guidance and mentorship have been instrumental in shaping both this work and my development as a researcher.

To Dr. Ludovica Griffanti, I owe immense appreciation for serving as an extraordinary research role model. Her thoughtful approach to tackling complex problems, her unwavering positivity in the face of challenges, and her remarkable ability to transform obstacles into opportunities have been a constant source of inspiration.

I am deeply grateful to Dr. Daniel Reich for sharing his exceptional breadth of knowledge in multiple sclerosis and MRI. From the outset, his openness to collaboration and generous exchange of ideas and insights fostered an environment conducive to rigorous scientific exploration and freedom to pursue diverse research directions.

It has been an extraordinary privilege to benefit from the guidance of Professor Heidi Johansen-Berg throughout this doctoral journey. Her accomplished, supportive leadership and steadfast mentorship have been a source of motivation and assurance.

Special acknowledgment is also due to Dr. Govind Nair, whose day-to-day supervision, mentorship, and genuine care for both the projects and the underlying science were invaluable. His guidance, constructive feedback, and attention to detail not only advanced the projects but also strengthened my research skills and shaped my growth as an independent scientist.

Deep gratitude is extended to the many collaborators who contributed to this research through their expertise and engagement in discussions on project ideas, results, and methodological approaches: in particular, Dr. Serhat Okar, Dr. María Inés Gaitán, Dr. Charidimos Tsagkas, and Dr. Tianxe Wu for their senior perspectives; Ashley Thommana and Henry Dieckhaus for their collaboration on projects; and all members of the TNS at NINDS, as well as the N-ACT and Plasticity groups at Oxford University, for fostering a supportive and stimulating research environment.

This doctoral work would not have been possible without the unwavering support of friends and family. Being surrounded by such amazing and caring people (too many to name individually, but each cherished) has been an extraordinary gift. Particular gratitude is owed to my parents, Trudy and Tim, whose endless belief in my abilities and readiness to help in countless ways formed a constant foundation. Whether meticulously proofreading chapters, listening patiently to practice presentations, or simply offering encouragement

during moments of frustration or breakthrough, their consistent support sustained every phase of this journey.

Deep gratitude is extended to all the participants in the National Institutes of Health studies, whose generous contributions of time and willingness to engage in research made this work possible.

Statement of Originality

I hereby certify that I am the sole author of this thesis, and that neither the whole nor any part of this thesis has been submitted for a degree or qualification at any other university or institution. This thesis is the result of my own work, except where specifically indicated in the text. Where the work is the outcome of a collaborative effort, I have acknowledged the contributions of each collaborator. A summary of my contributions and those of collaborators is outlined below:

1. Chapter 2.1 includes content based on collaborative work, limited to analyses involving *paired cohort 1* as outlined below. I completed all other analyses within the section.
    - Henry Dieckhaus (H.D.) independently implemented the transfer learning method for PLAn-7T, optimized model learning rates, and created comparison datasets (C-DEF 7T, C-DEF 3T, FreeSurfer).
    - H.D. also selected the case examples shown in Figures 2.2 and 2.3.
    - H.D. and I were jointly responsible for dataset curation, segmentation mask preparation, evaluation of model performance, and co-writing the related manuscript.
2. Statistical modelling in Chapter 3.4.3 was conducted in collaboration with Dr. Tianxe Wu (T.W.). For the clinical trial analyses (Chapters 3.2 and 3.3), random coefficient models and analysis of covariance models were performed using methods implemented by TW. In Chapter 3.1, I carried out the analyses based on methods outlined by TW.
3. Chapters 4.3.4 and 5.4 include contributions from Dr. Serhat Okar (S.V.O.) and Ashley Thomanaa (A.T.). A.T. extracted all clinical and demographic data. A.T. generated ultra-low field (ULF) and 3T segmentations for Chapter 4.1 and reviewed image processing outputs. S.V.O reviewed all reference segmentations and provided feedback on the ULF pipelines.
4. Neurological assessments and ratings were conducted by neurologists, whose initials are indicated in the respective methods sections.

- Dr. María I. Gaitán (M.I.G.) and Dr. Charidimos Tsagkas (C.T.) provided qualitative segmentation ratings for Chapter 2.1.
- Dr. Daniel S. Reich (D.S.R.) and M.I.G. provided visual assessments of lesion biomarkers (e.g., paramagnetic rim status) for Chapter 3.4.3.
- S.V.O. rated segmentation quality (Chapter 4.1), assessed changes in WMLs (Chapter 4.2), and provided radiological assessments of 3T WML burden (Chapter 5.4)
- C.T. and S.V.O. provided qualitative ratings of image quality in Chapter 4.3

5. All MRI datasets analysed in this thesis were acquired under protocols at the Translational Neuroradiology Section (TNS), National Institute of Neurological Disorders and Stroke (NINDS) (see Appendix A.1 for details). I participated in scanning a subset of participants. However, the majority of MRI data were collected and managed by TNS clinical staff and MRI radiographers.

To the best of my knowledge, this thesis does not infringe upon any copyrights or proprietary rights. All ideas, techniques, data, and materials from other sources have been fully acknowledged according to academic referencing standards. Permission to reuse published work and confirmation of authorship have been submitted to the university in a separate document and formally approved.

Abstract

Ultra-high field (UHF) and ultra-low field (ULF) magnetic resonance imaging (MRI) offer complementary approaches for studying and monitoring white matter lesions (WMLs) in multiple sclerosis (MS). UHF MRI provides enhanced sensitivity to chronic lesion pathology, while ULF MRI enables portable and accessible imaging. Existing image analysis tools, developed primarily for conventional MRI, often fail at these field strengths. This thesis develops and applies dedicated image analysis tools for WMLs at UHF (7T) and ULF (64mT), addressing field-specific research questions.

In the first part of this thesis, 7T tools were developed to quantify longitudinal changes in chronic lesions. Three analytical tools were developed: a 3T–7T transfer learning model for whole-brain and WML segmentation; an automated pipeline for tracking lesion quantitative T1 and volume changes; and a web-based application for standardized radiological lesion assessment. Application to a large longitudinal natural history cohort and two clinical trials revealed new insights into the long-term evolution of chronic lesions with and without paramagnetic rims, while establishing important methodological considerations for implementing 7T-derived biomarkers in treatment monitoring.

The second part addresses ULF MRI tool development to determine whether ULF imaging can capture clinically relevant lesion metrics compared to a 3T clinical standard. This work involved adapting transfer learning models for ULF WML segmentation, implementing a longitudinal subtraction pipeline to support radiological detection of evolving lesions, and evaluating a Fourier-based super-resolution approach combining orthogonal acquisitions to improve image quality. Analysis of paired 3T–64mT data demonstrated that 3T-derived lesion metrics and their respective associations with clinical disability could be captured at ULF. These findings contribute to the characterization of ULF's sensitivity and detection limits for WML, thereby informing its potential clinical utility.

Together, this research demonstrates that leveraging UHF and ULF can deepen our understanding of lesion pathophysiology and extend the reach of WML imaging in MS.

Keywords: Multiple sclerosis, MRI, ultra-low field, ultra-high field, white matter lesions, paramagnetic rim lesions

Contents

List of Figures **vii**

List of Tables **xiii**

1 Introduction and Background **1**

1.1 Motivation and Objectives 1
1.2 Overview of Multiple Sclerosis 5
1.3 Conventional MRI of White Matter Lesions in Multiple Sclerosis 9
1.4 Advanced MRI of White Matter Lesions in Multiple Sclerosis . . 11
1.4.1 7T MRI for Chronic White Matter Lesions Assessment . 12
1.4.2 Ultra-Low Field MRI for White Matter Lesion Assessment 20
1.5 Challenges in Studying White Matter Lesions with Advanced MRI 22
1.5.1 Quantification of White Matter Lesions 22
1.5.2 Tracking Longitudinal Changes in White Matter Lesions 24
1.6 Summary of Research Objectives 28

2 Development of 7T MRI Analysis Tools for Longitudinal Tracking of Multiple Sclerosis Lesions **29**

2.1 Development of 7T Whole Brain and Lesion Segmentation from a Single MRI Modality 30
2.1.1 Introduction 30
2.1.2 Methods 33
2.1.3 Results 37
2.1.4 Discussion 43
2.2 Impact of Longitudinal Analysis Pipelines Quantitative T1 and Volume 47
2.2.1 Introduction 47
2.2.2 Methods 48
2.2.3 Results 54

2.2.4 Discussion 57
2.3 Development of a Visualization Application to Streamline Multimodal Longitudinal Lesion Assessment 64
2.3.1 Introduction 64
2.3.2 Methods 66
2.3.3 Evaluation and Feedback 73

3 Longitudinal Assessment of Quantitative T1 and Volume in Chronic White Matter Lesions at 7T 77
3.1 Longitudinal Changes of Quantitative T1 and Volume in Chronic White Matter Lesions in a 7T Natural History Dataset 79
3.1.1 Introduction 79
3.1.2 Methods 82
3.1.3 Results 88
3.1.4 Discussion 96
3.2 Effect of Anakinra on Chronically Inflamed White Matter Lesions in Multiple Sclerosis 101
3.2.1 Overview of Interleukin-1β and Anakinra in Multiple Sclerosis 101
3.2.2 Methods 103
3.2.3 Results 113
3.2.4 Conclusions 123
3.3 Effect of Tolebrutinib on Chronically Inflamed White Matter Lesions in Multiple Sclerosis 124
3.3.1 Overview of Bruton's tyrosine kinase and Tolebrutinib in MS 125
3.3.2 Methods 128
3.3.3 Results 133
3.3.4 Conclusions 143
3.4 Discussion of MRI Results from Clinical Trials 145
3.4.1 Radiological Assessment of the Paramagnetic Rim 145
3.4.2 Quantitative White Matter Lesion Analysis 147
3.4.3 Conclusions 149

4 Development of Ultra-Low Field MRI Analysis Tools for Tracking of Multiple Sclerosis Lesions 151

4.1 Adaptation of 7T PLAn for Ultra-Low Field Lesion Segmentation 152

4.1.1 Introduction . . . 152

4.1.2 Methods . . . 154

4.1.3 Results . . . 163

4.1.4 Discussion . . . 177

4.2 Subtraction Imaging Pipeline to Assess Longitudinal Lesion Changes in Multiple Sclerosis at Ultra-Low-Field . . . 182

4.2.1 Introduction . . . 182

4.2.2 Methods . . . 184

4.2.3 Results . . . 190

4.2.4 Discussion . . . 196

4.3 Improving Ultra-Low-Field MRI Resolution using Partially Acquired K-Space . . . 199

4.3.1 Introduction . . . 199

4.3.2 Methods . . . 202

4.3.3 Results . . . 207

4.3.4 Discussion . . . 212

5 Preliminary Application of Ultra-Low Field MRI for Tracking Multiple Sclerosis White Matter Lesions 217

5.1 Introduction . . . 217

5.2 Methods . . . 218

5.2.1 Clinical Correlates of Cross-sectional Total Lesion Volume 218

5.2.2 Clinical Correlates of Longitudinal Lesion Evolution . . . 219

5.3 Results . . . 221

5.3.1 Clinical Correlates of Cross-sectional Total Lesion Volume 221

5.3.2 Clinical Correlates of Longitudinal Lesion Evolution . . . 225

5.4 Discussion . . . 228

6 Conclusions and Future Work 231

6.1 7T Tools and Applications . . . 231

6.2 Ultra-Low Field MRI Tools and Applications . . . 237

6.3 Concluding Thoughts on Ultra-High and Ultra-Low Field MRI in Multiple Sclerosis . . . 239

Appendices

A Dataset Details **243**
A.1 IRB Protocols . . . 243
A.2 Datasets used in Chapter 2.3.3 . . . 244
A.3 Datasets used in Chapter 3.4.3 . . . 245
A.3.1 Dataset used in Chapter 3.1 . . . 246
A.3.2 Dataset used in Chapter 3.2 . . . 247
A.3.3 Dataset used in Chapter 3.3 . . . 247
A.4 Datasets used in Chapter 4.3.4 . . . 247
A.4.1 Dataset used in Chapter 4.1 . . . 247
A.4.2 Dataset used in Chapter 4.2 . . . 249
A.4.3 Dataset used in Chapter 4.3 . . . 250
A.5 Datasets used in Chapter 5.4 . . . 251

B Supplementary Data for Chapter 2.3.3 **253**
B.1 Supplementary Data for Chapter 2.1 . . . 253
B.2 Supplementary Data for Chapter 2.2 . . . 254

C Supplementary data for Chapter 3.4.3 . **257**
C.1 Supplementary data for Chapter 3.1 . . . 257
C.1.1 Summary of Covariates Considered for Linear Mixed Models . . . 257
C.1.2 Sensitivity Analyses for LMMs in Chapter 3.1 . . . 259
C.2 Supplementary Data for Chapter 3.2 . . . 265
C.3 Supplementary Data for Chapter 3.3 . . . 269
C.3.1 Statistical analyses for BRaKE-MS . . . 272
C.3.2 Exploratory analysis of image processing pipeline on lesion qT1 and volume. . . . 275

D Supplementary data for Chapter 4.3.4. **281**
D.1 Supplementary data for Chapter 4.1 . . . 281
D.1.1 Pairwise Statistics for Segmentation Overlap Metrics . . 281
D.1.2 Performance of Automatic Segmentation Algorithm across ULF Acquisition Parameters . . . 298
D.2 Supplementary data for Chapter 4.2 . . . 300

D.2.1 Normalisation during ULF SUBLONG Pipeline Development 300
D.2.2 Impact of ULF Acquisition Parameters on ULF SUBLONG Analysis 300
D.3 Supplementary data for Chapter 4.3 306
D.3.1 Effect of FouSR Sampling Size on Quantitative Evaluation Metrics 306

E Supplementary data for Chapter 5.4 **309**
E.1 Extended Cross-Sectional Analyses to All Automatic ULF Segmentation Algorithms 309

References **321**

List of Figures

1.1 Clinical progression of multiple sclerosis (MS) across different disease courses. 7
1.2 Recommended sequence contrasts for routine clinical scans for MS at 3T 10
1.3 Chronic lesions on T2*-GRE magnitude and phase sequence at 7T. 15
1.4 Example chronic lesions on 7T MP2RAGE contrasts at 0.5mm isotropic resolution. 20

2.1 Overview of Datasets, PLAn-7T Algorithm and Evaluations of 7T segmentation methods 35
2.2 Representative Examples of Whole Brain Segmentations at 7T. . 39
2.3 Representative Examples of Segmentation Outputs at 7T. . . . 40
2.4 Quantitative evaluation of 7T segmentation algorithms compared to 3T Reference 41
2.5 Evaluation of PLAn 7T and nnU-Net 7T WML segmentation on an unseen cohort. 42
2.6 Reproducibility Evaluation of PLAn 7T and nnU-Net 7T Segmentations 43
2.7 Schematic of pipelines for longitudinal test–retest analysis . . . 51
2.8 Examples of Lesion Characteristics for Test-Rest Analysis 53
2.9 Within-scanner Test-Retest Variability of MP2RAGE qT1 Times in GM, WM, and CSF 55
2.10 Lesion Analysis Stage 1: Within-scanner Test-Retest Variability of MP2RAGE qT1 times and volume 56
2.11 Lesion Analysis Stage 1: WML characteristics and their association with test-retest variability in median qT1 and lesion volume 57

2.12 Lesion Analysis Stage 2: Within-scanner Test-Retest Variability of MP2RAGE qT1 times and volume 58
2.13 Example of a current method for visually rating lesions. 66
2.14 Overview of Flask Visualization Web Page Application 68
2.15 Overview of the steps involved in developing the Flask application 72

3.1 Overview of image processing steps for quantitative analysis of PRL+ vs PRL-. 86
3.2 Longitudinal MRI scan timepoints for all study participants (n = 81) 90
3.3 PRL+ vs. PRL– Lesions at Baseline and Longitudinally. 92
3.4 PRL+ vs. PRL- lesions by lesion location and layer at baseline and longitudinally. 95
3.5 Overview of ATaC study design 104
3.6 Overview of image processing steps for Clinical Trial 107
3.7 Examples of Flask applications used for ATaC-MS Longitudinal PRL Ratings 110
3.8 ATaC-MS PRL+ with reduced rim conspicuity between Week 0 – Week 12. 119
3.9 Lesion qT1 and volume in ATaC-MS trial completers 121
3.10 Longitudinal Changes in Clinical Scores in ATaC-MS Participants 123
3.11 Overview of BRaKE-MS study design 129
3.12 The PRL+ from a participant in treatment cohort A, with a rim that became less conspicuous over time 137
3.13 Baseline median qT1 and volume in PRL+ and PRL- in BRaKE-MS 138
3.14 Longitudinal changes in qT1 and lesion volume in PRL+ and PRL- lesions in BRaKE-MS, estimated using RCM 140
3.15 qT1 and volume trajectories in PRL+ and PRL- in BRaKE-MS, as measured by RRME model 142
3.16 Longitudinal Changes in Clinical Scores in BRaKE-MS 143

4.1 Comparison of 3T and ULF Manual Lesion Segmentations. 164
4.2 Visual detection of individual WMLs on ULF. 166
4.3 Comparison of total WML volume across automatic segmentation methods to reference WML masks. 168

4.4 Evaluation of automatic WML segmentation performance in original ULF resolution. 170
4.5 Evaluation of automatic WML segmentation performance in original ULF resolution. 172
4.6 Example WML masks for ULF Aided Reference and automatic methods overlaid on ULF FLAIR. 174
4.7 Effects of probability mask thresholding on DSC for automatic ULF WML segmentation 176
4.8 Comparison of overlap metrics across different models, using optimized probability thresholds for MIMoSA, nnU-Net and PLAn models. 177
4.9 Overview of study design 186
4.10 Radiological definitions of changing lesions (CL) including enlarging (EL) and shrinking lesions (SL) and new lesions (NL) using preprocessed and SUBLONG images 188
4.11 Examples of CLs visible on both 3T and ULF across preprocessed and SUBLONG images. 190
4.12 Quantitative analysis of radiologically identified lesion dynamics 193
4.13 Comparison of 3T and ULF volume measures for EL, SL and NL radiologically detected on ULF. 195
4.14 An example of how sampling MRI in the frequency domain can affect image quality 201
4.15 Overview of image pre-processing and image combination for FouSR. 204
4.16 Examples of image quality across different methods in a standard Hyperfine phantom. 208
4.17 Intensity plots of ULF Axial, ULF Coronal, ULF Average, and FouSR. 208
4.18 Example of different methods applied to a participant with high lesion load. 209
4.19 Comparison of quantitative scan quality metrics in in-vivo data ($n = 10$) across methods. 211
4.20 Comparison of qualitative ratings for different methods. 212

5.1 Total ULF and 3T WML volumes across radiologically defined lesion burden categories 221

5.2 Clinical correlations with total WML volume at 3T and ULF. . 224
5.3 Spearman correlation matrix showing associations between change in clinical measures (Δ) and radiological lesion activity measures from 3T and ULF SUBLONG MRI. 226
5.4 Longitudinal clinical score changes grouped by case-level ULF lesion activity detection. 228

A.1 Example of T2*-GRE Magnitude and Phase acquisitions 246

B.1 Evaluation of PLAn-7T and nnU-Net 7T WML segmentation on an unseen cohort 254

C.1 An overview of PRL+ lesions included in various analyses in the ATaC-MS clinical trial 266
C.2 An overview of PRL+ lesions included in various analyses in the BRaKE-MS clinical trial 271
C.3 Inverse relationship between longitudinal lesion qT1 and volume in PRL- lesions 276
C.4 Comparison of qT1 trajectories using different lesion segmentation masks. 278
C.5 Comparison of Jacobian determinant (JD) from participant-wise affine registration with changes in lesion volume and qT1 over time 280

D.1 Evaluation of automatic WML segmentation performance in $1 \times 1 \times 1$ mm^3 resolution 296
D.2 Evaluation of automatic WML segmentation performance in $1 \times 1 \times 1$ mm^3 resolution (subset) 297
D.3 Performance of Automatic Segmentation Algorithm across ULF Acquisition Parameters 299
D.4 Comparison of normalisation methods during ULF SUBLONG pipeline development 300
D.5 Participant-level WML detection of CL by ULF acquisition related metrics 302
D.6 Participant-level WML detection of NL by ULF acquisition related metrics 303
D.7 Impact of ULF acquisition parameters on the individual CLs) radiologically detected vs missed at ULFs 304

D.8 Impact of ULF acquisition parameters on the individual NLs) radiologically detected vs missed at ULFs 304
D.9 Difference between 3T and ULF measured longitudinal volume change for NLs by ULF acquisition-related metrics) radiologically detected vs missed at ULFs. 305
D.10 Difference between 3T and ULF measured longitudinal volume change for NLs by ULF acquisition-related metrics) radiologically detected vs missed at ULFs (outlier removed) 306
D.11 Difference between 3T and ULF NL volume by ULF follow-up image acquisition parameters 306
D.12 The effect of varying the centre matrix size in FouSR sparse sampling on quantitative image metrics. 307

E.1 Extended Analysis of Total Lesion Volume by Radiological Lesion Burden 311
E.2 Extended Correlations Analysis between Total Lesion Volume and Clinical Measures 313

List of Tables

2.1 Qualitative Rating Definitions for 7T Segmentation Evaluations. 36
2.2 Qualitative Ratings of 7T Segmentations. 38
2.3 Volumetric Reproducibility Evaluation of nnU-Net 7T and PLAn 7T Segmentation 44

3.1 Natural History Linear Mixed Model Designs 87
3.2 Demographics and baseline characteristics of all participants . . 89
3.3 Participant demographics, whole brain MRI characteristics, lesion-level MRI characteristics, and clinical scores for completers and non-completers. Data shown as mean ± SD, with ranges [min, max] where applicable. 114
3.4 Inter- and Intra-rater reliability for PRL ratings in ATaC-MS . . 116
3.5 ATaC-MS Longitudinal Rim Changes 118
3.6 Participant demographics, whole brain MRI characteristics, lesion-level MRI characteristics, and clinical scores for Non-Treatment and Treatment Cohort. Data shown as mean ± SD, with ranges [min, max] where applicable. 134
3.7 Inter- and Intra-rater reliability for PRL ratings in BRaKE-MS 135
3.8 BRaKE-MS Longitudinal Rim Changes 136

4.1 Participant Characteristics and Descriptive Statistics in Training and Validation Cohorts. Data shown as count or mean ± SE, with ranges [min, max] where applicable. 156
4.2 Overview of Automatic Segmentation Methods 157
4.3 Qualitative Rating Scale Definitions for Automatic Segmentation Methods at 64mT 161
4.4 Overview of all ULF WML segmentation volume comparisons . 165
4.5 Qualitative Ratings of Automatic WML Segmentations at ULF 173
4.6 Longitudinal Lesional Change at Case and Lesion Level 192

4.7 Description of qualitative rating scales for WML and Sulci and Gyri 206

5.1 Sub-cohorts of participants with available clinical scores. Numeric variables are shown as mean ± standard error (SE) with range [min, max] where applicable. Counts are shown for categorical variables. 222
5.2 Clinical characteristics of MS participants with longitudinal radiological lesion activity assessments, grouped by availability of longitudinal clinical scores. Numeric variables are shown as mean ± standard error (SE) with range [min, max] where applicable. Counts are shown for categorical variables. 225

A.1 Institutional Review Board (IRB) Protocols for MRI Datasets. . 243
A.2 MRI acquisition parameters for datasets in Chapter 2.3.3 244
A.3 Summary of Participant Demographics for Dataset used in Chapter 2.3.3 245
A.4 7T MRI acquisition parameters for datasets in Chapter 2.3.3 . . 246
A.5 MRI acquisition parameters used in Chapter 4.1 248
A.6 ULF MRI acquisition used in Chapter 4.1 by group 249
A.7 MRI acquisition parameters used in Chapter 4.2 250
A.8 Summary of Participant Characteristics for Dataset used in Chapter 4.2 250
A.9 MRI acquisition parameters used in Chapter 4.3 251
A.10 Summary of Participant Characteristics for Dataset used in Chapter 4.3 251
A.11 MRI acquisition parameters used in cross-sectional analyses in Chapter 5.4 252
A.12 MRI acquisition parameters used in Longitudinal Analysis in Chapter 5.4 252

B.1 Bias and limits of agreement across pipelines 255
B.2 Bland-Altman Statistics for Lesion Analysis Stages 1 and 2 . . . 256

C.1 Sensitivity analysis 2 for baseline qT1 and volume LMM: PRL+ vs PRL- 260
C.2 Sensitivity analysis 1 for longitudinal qT1 and volume LMM: PRL+ vs PRL- 260

C.3 Sensitivity analysis 2 for longitudinal qT1 and volume LMM: PRL+ vs PRL- 261
C.4 Sensitivity analysis 2 for baseline qT1 and volume LMM contrasts by PRL and lesion location 262
C.5 Sensitivity analysis 1 for longitudinal qT1 and volume LMM contrasts by PRL and lesion location 263
C.6 Sensitivity analysis 2 for longitudinal qT1 and volume LMM contrasts by PRL and lesion location 264
C.7 Inclusion and Exclusion Criteria for ATaC-MS Clinical Trial . . 265
C.8 Post-hoc Pairwise Comparisons from RM-ANCOVA of Normalized Lesion qT1 (Full Dataset with Imputed Missing Data) . . . 266
C.9 Post-hoc Pairwise Comparisons from RM-ANCOVA of Normalized Lesion qT1 after removing statistical outliers, 8 timepoints (Full Dataset with Imputed Missing Data) 267
C.10 Post-hoc Pairwise Comparisons from RM-ANCOVA of Normalized Lesion Volume (Full Dataset with Imputed Missing Data) . 267
C.11 Post-hoc Pairwise Comparisons from RM-ANCOVA of Normalized Lesion qT1 (Excluding participants with Imputed Data) . . 268
C.12 Post-hoc Pairwise Comparisons from RM-ANCOVA of Normalized Lesion volume (Excluding participants with Imputed Data) 268
C.13 Inclusion and Exclusion Criteria for BRaKE-MS Clinical Trial Treatment Cohort 269
C.14 Inclusion and Exclusion Criteria for BRaKE-MS Clinical Trial Non-Treatment Cohort 270
C.15 BRaKE-MS RCM estimated slopes after removing statistical outliers 272
C.16 BRaKE-MS RRME details for pairwise post hoc comparisons in BRaKE-MS analysis 273
C.17 BRaKE-MS Pairwise Post-Hoc RRME Model 273
C.18 BRaKE-MS Pairwise Post-Hoc RRME Model Treatment Cohort 274
C.19 BRaKE-MS Pairwise Post Hoc Comparisons for volume in the Treatment Cohort (RRME Model) 275

D.1 DSC pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods 282

D.2 FPR pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods 283
D.3 FNR pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods 284
D.4 FPR cluster pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods 285
D.5 FNR cluster pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods 286
D.6 DER pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods 287
D.7 OER pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods 288
D.8 DSC pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods (ULF Blinded Subset) 289
D.9 FPR pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods (ULF Blinded Subset) 290
D.10 FNR pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods (ULF Blinded Subset) 291
D.11 FPR Cluster pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods (ULF Blinded Subset) 292
D.12 FNR cluster pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods (ULF Blinded Subset) 293
D.13 DER pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods (ULF Blinded Subset) 294
D.14 OER pairwise post-hoc comparisons of ULF automatic WML overlap metrics across segmentation methods (ULF Blinded Subset) 295
D.15 Table Heterogeneity in ULF Acquisition Parameters Between Baseline and Follow-Up 301

E.1 Statistically Summary of Total WML Volume across Radiologically-defined WML Burden Categories 312
E.2 Overview of linear regression between total WML volume and age across segmentation methods 314
E.3 Overview of age-adjusted linear regression between total WML volume and EDSS across segmentation methods 315
E.4 Overview of age-adjusted linear regression between total WML volume and Scripps across segmentation methods 316
E.5 Overview of age-adjusted linear regression between total WML volume and SDMT across segmentation methods 317
E.6 Overview of age-adjusted linear regression between total WML volume and 9HPT across segmentation methods 318
E.7 Overview of age-adjusted linear regression between total WML volume and PASAT across segmentation methods 319

Chapter 1

Introduction and Background

Contents

1.1 Motivation and Objectives 1

1.2 Overview of Multiple Sclerosis 5

1.3 Conventional MRI of White Matter Lesions in Multiple Sclerosis 9

1.4 Advanced MRI of White Matter Lesions in Multiple Sclerosis 11

1.5 Challenges in Studying White Matter Lesions with Advanced MRI 22

1.6 Summary of Research Objectives 28

1.1 Motivation and Objectives

Multiple sclerosis (MS) is a chronic immune-mediated disease of the central nervous system (CNS). The typical clinical course begins with a relapsing phase marked by acute inflammatory episodes and clinical relapses, followed by a progressive phase characterized by less frequent relapses but gradual accumulation of disability over time (Frischer et al., 2009; Lassmann, 2013; Lublin et al., 2014). Although clinical relapses can be effectively managed with disease-modifying therapies (DMT) that target acute inflammation, many patients continue to accumulate disability, highlighting a therapeutic gap in addressing progressive aspects of MS (Cree et al., 2019; Hauser and Cree, 2020). Pathologically, the evolution from the relapsing to progressive phases of MS is thought to reflect a shift from acute focal inflammatory injury to more diffuse chronic inflammation and neurodegeneration (Lassmann, 2017).

A hallmark of MS is the presence of focal demyelinating lesions, particularly within the white matter (WM) of the brain. These white matter lesions (WMLs) are commonly visualized using magnetic resonance imaging (MRI), a critical tool in both clinical management and MS research (Wattjes et al., 2021). The value of MRI in the context of MS depends greatly on the analytical tools used to extract quantitative measures from MRI. Particular analytical challenges arise when non-standard MRI approaches are used. This thesis considers those analytical challenges, with the aim of enhancing the value of non-standard MRI for clinical trials and clinical management of MS.

In this work, conventional MRI (cMRI) refers to clinically standard pulse sequences, including T1-weighted, gadolinium-enhanced T1-weighted, T2-weighted, and fluid-attenuated inversion recovery (FLAIR) acquired at high field strengths (HF) of 1.5T or 3T. These sequences are effective for detecting new and enlarging WML, as well as acute inflammation associated with MS lesion formation (Wattjes et al., 2021). Importantly, the availability of open-source and well-established image processing tools has facilitated the detection of lesion-related disease activity on cMRI, enabling it to be adopted as a key endpoint in clinical trials.

Although cMRI is central to clinical practice, recent developments in advanced MRI methods, including new sequences and non-standard field strengths, offer complementary advantages for assessing lesion-related disease activity. Ultra-high field (UHF) MRI, typically at 7 tesla (T) for in-vivo clinical imaging, provides high spatial resolution and improved sensitivity for characterizing lesion pathology, which is valuable for studying mechanisms underlying progressive MS (Harrison et al., 2024). In contrast, ultra-low field (ULF) MRI offers portable and accessible imaging, with the potential to capture lesion-relevant disease activity in settings where conventional imaging is limited or unavailable. This thesis considers opportunities and challenges for both UHF and ULF MRI to provide added value in the context of MS trials and management.

7T MRI offers key advantages in studying prominent mechanisms that contribute to progression, such as demyelination and chronic inflammation in WML (Harrison et al., 2024; Kolb et al., 2022). On susceptibility-sensitive imaging (SSI), a paramagnetic rim is detectable around a subset of chronic lesions, which co-localizes with iron-laden, pro-inflammatory microglia and macrophages accumulating at the lesion border (Dal-Bianco et al., 2017; Yao et al., 2012). These lesions, known as paramagnetic rim lesions (PRLs), are

considered imaging biomarkers of chronic, low-grade inflammation thought to contribute to progressive MS. PRLs are typically distinguished from chronic lesions without a rim (referred to as PRL+ and PRL-, respectively) and are associated with more destructive pathology and greater disability (Absinta et al., 2019, 2016). The sub-millimetre resolution achievable at 7T increases the sensitivity to detecting PRLs and allows for the monitoring of changes within the rim over time.

Additionally, myelin-sensitive imaging methods such as quantitative T1 (qT1) mapping offer insights into lesion microstructure (Granziera et al., 2021). At 7T, qT1 mapping benefits from improved contrast and signal separation, which can reveal heterogeneity both between and within lesions. While PRLs have been linked to poor clinical outcomes, their long-term evolution is still uncertain. 7T qT1 mapping is sensitive to both demyelination and remyelination, highlighting its potential as a biomarker for clinical trials (Kolb et al., 2021). Furthermore, the paramagnetic rim itself has been shown to disappear in some lesions over time, with recent studies reporting a resolution of the paramagnetic rim after a median of approximately seven years (Absinta et al., 2021; Dal-Bianco et al., 2021). Thus, chronic inflammation may be reversible, and PRLs could be a target for future treatments.

At the other extreme of MRI field strength, portable ULF MRI offers an emerging opportunity to increase access to imaging for patients who face logistical or clinical barriers to cMRI (Arnold et al., 2023; Deoni, Medeiros, Deoni, Burton, Beauchemin, D'Sa, Boskamp, By, McNulty, Mileski, Welch and Huentelman, 2022; Wald et al., 2020). These systems can be low-cost, mobile, and provide point-of-care use in settings where cMRI is not possible. However, ULF MRI is a relatively new technology, and its role in MS is not yet well established (Arnold et al., 2023). While ULF MRI lacks the spatial and contrast resolution of cMRI systems, it may serve as a useful supplement to cMRI for follow-up imaging or screening in MS. An important area of investigation is to determine which aspects of MS pathology, such as WML, are visible with ULF MRI, and whether ULF-derived measurements align with those from cMRI. Understanding these limitations and capabilities is essential before ULF MRI can be meaningfully integrated into clinical workflows for MS care and WML monitoring.

Although 7T and ULF MRI are used to address different research and clinical questions in MS, both share a common challenge: existing MRI analysis

tools are typically developed and optimized for cMRI and do not always translate well to non-standard field strengths. At 7T, increased field inhomogeneities and stronger bias fields can compromise the performance of standard image processing tools, such as segmentation and registration algorithms (Harrison et al., 2024). Similarly, there remains a lack of image processing tools tailored to handle ULF MRI's limitations in resolution and contrast (Arnold et al., 2023). A core aim of this thesis is to develop, adapt, and validate image processing tools to reliably track lesions and answer modality-specific research questions.

In this thesis, I address key methodological challenges in the analysis of MS lesions using non-standard MRI field strengths. I focus on the development and adaptation of image processing tools for 7T and ULF (0.064T) MRI to support the investigation of WML in MS. At 7T, this research centres on developing methods to reliably track the evolution of qT1 and volume of chronic lesions and evaluating whether the chronic lesions subset, PRLs, can be modulated by experimental treatment. At ULF, I examine how existing tools can be adapted for WML analysis to determine whether ULF MRI can capture WML characteristics observed at 3T.

In the remaining sections of Chapter 1.6, I introduce the clinical and pathological features of MS, with an emphasis on WMLs and their assessment using MRI. I review conventional and advanced MRI approaches for WML evaluation. At 7T, I focus on SSI to identify PRLs and qT1 mapping to assess tissue integrity, as these contrasts are used in this thesis for chronic lesion analysis. At ULF, I focus on the potential of ULF MRI in MS to provide context for the experimental work in this thesis that explores ULF MRI's ability to detect and follow WMLs. I then conclude the chapter by outlining key challenges in WML quantification in general and across different field strengths. These sections incorporate text and figures from a book chapter under review (Advanced MRI in Multiple Sclerosis, Donnay CA et al., Elsevier).

In Chapter 2.3.3, I develop tools for three components of the image processing pipeline needed to assess longitudinal changes in qT1 and volume in chronic lesions at 7T. This includes: (i) training a whole-brain and WML segmentation model using 7T MP2RAGE images; (ii) comparing longitudinal processing pipelines for tracking lesion-specific qT1 and volume; and (iii) building a visualization application to support radiological evaluation of PRLs.

In Chapter 3.4.3, I apply these tools to investigate how lesion qT1 and volume evolve over time in a 7T natural history dataset, and whether these

features, along with rim characteristics, are modulated by investigational treatments in two clinical trials with anakinra or tolebrutinib.

In Chapter 4.3.4, I adapt analysis tools for the study of WML at ULF. This includes: (i) adapting the 7T segmentation model for ULF WML segmentation; (ii) developing a subtraction-based pipeline to assess radiological lesion changes; and (iii) enhancing spatial resolution by combining orthogonal ULF acquisitions.

In Chapter 5.4, I present preliminary results evaluating whether the ULF tools can capture clinically relevant WML information observed on 3T MRI using a unique dataset of paired 3T and ULF scans on the same patients.

Finally, in Chapter 6, I summarize the main contributions of this thesis and discuss future directions.

1.2 Overview of Multiple Sclerosis

MS, a chronic inflammatory and neurodegenerative disease of the CNS, is a leading cause of non-traumatic neurological disability in young adults, with a typical age of clinical onset in the early thirties. Females are twice as likely to have MS compared to males (Walton et al., 2020). The clinical presentation of MS is heterogeneous, and typical symptoms may include fatigue, muscle weakness, sensory disturbances, impaired coordination, visual changes, or cognitive dysfunction (Macias Islas and Ciampi, 2019).

To monitor disease severity and progression, several clinical outcome measures are routinely used in both research and clinical settings (Inojosa et al., 2020). The Expanded Disability Status Scale (EDSS) is the most widely used tool for quantifying physical disability in MS, with a primary focus on mobility (Kurtzke, 1983). However, it has limited sensitivity when it comes to cognitive impairment and subtle changes in disability (Weinstock-Guttman et al., 2022). To capture a broader range of disabilities, additional functional tests are often incorporated (Fischer et al., 1999). The Timed 25-Foot walk (25FT) evaluates lower extremity function and gait speed, while the 9-hole peg test (9HPT) assesses upper extremity dexterity. The Scripps Neurological Rating Scale (Scripps) offers a composite measure of neurological function and has been utilized in various clinical trials to complement EDSS scoring (Sipe et al., 1984). Cognitive performance is typically assessed using tests such as the Paced Auditory Serial Addition Test (PASAT) and the Symbol Digit Modalities Test

(SDMT). Both tests are sensitive to impairments in information processing speed and working memory, which are frequently affected in MS. Together, these assessments provide valuable context for understanding the clinical impact of the underlying pathological processes and are often used as secondary endpoints in clinical trials. In this thesis, these measures are included as secondary endpoints in the 7T clinical trials and are used to correlate clinical outcomes with ULF-derived MRI metrics.

Traditionally, MS has been categorized into distinct clinical phenotypes, with the most common initial presentation being relapsing-remitting MS (RRMS), characterized by episodes of neurological worsening (relapses) followed by periods of partial or complete recovery (remissions). Over time, most RRMS patients transition to secondary progressive MS (SPMS), characterized by gradual and irreversible accumulation of disability independent of relapses (Frischer et al., 2009; Lassmann, 2013; Lublin et al., 2014). In approximately 10-15% of individuals with MS, the disease presents as a steady progression of neurological decline from onset, without distinct clinical relapses, known as primary progressive MS (PPMS) (Montalban et al., 2017). While clinical phenotypes are reported, analyses in this thesis are not stratified by subtype, as these classifications are not strictly rooted in underlying biology and can overlap substantially. Instead, more quantifiable indicators of disease course are considered. For example, in clinical trials targeting progressive aspects of MS (chronic inflammation), measures such as *time since last relapse* or *time since the most recent focal inflammatory event* are used, as they provide more direct and reproducible markers of whether participants are in a progressive-dominant phase of the disease.

In addition to the main clinical subtypes of MS, clinically isolated syndrome (CIS) and radiologically isolated syndrome (RIS) are common descriptors related to early or prodromal stages of disease expression. CIS typically refers to the first neurological event caused by CNS inflammation and demyelination (Miller et al., 2012). Based on this definition, CIS does not necessarily indicate MS; for example, isolated optic neuritis or transverse myelitis may represent non-MS CIS if diagnostic criteria for MS are not met. However, in individuals who go on to meet the diagnostic criteria for MS, the CIS event is effectively the first clinical relapse of RRMS. Accordingly, for the experimental cohorts in this thesis, all MS participants with a history of CIS are classified under the RRMS category, as they represent different points along the same disease trajectory.

It is also possible that incidental MRI findings consistent with demyelination may precede any clinical symptoms. These findings, known as RIS, may represent a prodromal stage of the disease in individuals who are later diagnosed with MS (Lublin et al., 2014).

Although these classifications of clinical subtypes have historically helped to stratify patients for clinical and research purposes, there is increasing recognition that MS exists on a clinical and pathological continuum (Figure 1.1). MS reflects a dynamic spectrum in which inflammation, demyelination, and neurodegeneration co-occur throughout the disease course (Hauser and Cree, 2020). While both inflammation and neurodegeneration are already present at disease onset, their relative contributions shift over time with relapsing forms typically dominated by inflammatory activity, and progressive phases more strongly associated with neurodegeneration.

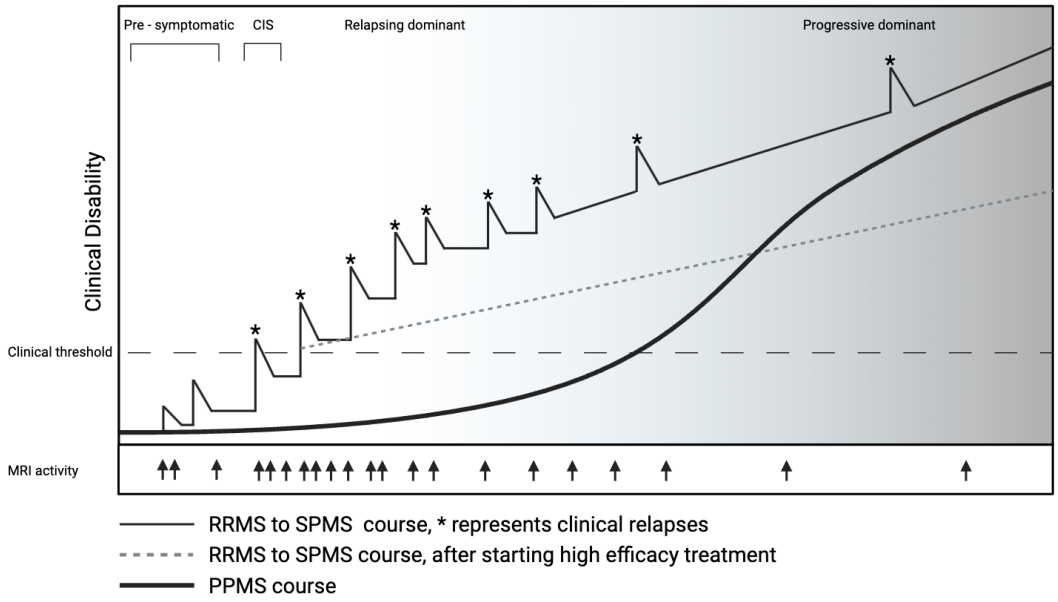


Figure 1.1: Clinical progression of multiple sclerosis (MS) across different disease courses. RRMS progressing to SPMS is shown as a thin solid line, and PPMS is shown as a thick solid line. In RRMS transitioning to SPMS, patients undergo clinical relapses denoted by asterisks (*) during the relapsing-dominant phase. Despite this, progression is noted during both the RRMS and SPMS phases. As the disease moves into the progressive-dominant stage, the rate of clinical relapses diminishes. The introduction of highly effective DMTs for RRMS has significantly decreased the frequency of clinical relapses, depicted by the dotted grey line. The clinical threshold, depicted by the black dotted line, shows the level at which subclinical disease activity (inflammatory relapses) starts to cause clinical symptoms that a person can perceive. This figure is reproduced from Donnay CA et al., Advanced MRI in Multiple Sclerosis, Elsevier, book chapter under review.

The pathogenesis of MS involves a complex interplay between peripheral immune activation and CNS-resident pathological processes. In the early relapsing phase, adaptive immunity predominates. Autoreactive CD4+ T cells, B

cells, and monocytes infiltrate the CNS via a compromised blood-brain barrier (BBB) ((Afzali and Korn, 2025; Lassmann, 2018*a*)). These immune cells drive focal inflammation, causing demyelination, oligodendrocyte loss, and axonal injury (Frischer et al., 2009; Hauser and Cree, 2020). This cascade of damage leads to the formation of focal demyelinating lesions, typically centred around small parenchymal veins.

As MS progresses, particularly in SPMS and PPMS, adaptive immune responses wane, while chronic, compartmentalized inflammation within the CNS becomes more prominent. This phase is characterized by innate immune mechanisms involving activated microglia and astrocytes, as well as iron accumulation, oxidative stress, and mitochondrial dysfunction, all of which contribute to a neurodegenerative environment (Kuhlmann et al., 2023). In addition to these mechanisms of progression, the brain's capacity for repair and compensation gradually declines (Kuhlmann et al., 2023). Over time, an age-related decline in neural reserve and repair processes, such as remyelination and neuroplasticity, reduces resilience to injury, facilitating clinical progression. This shift from focal acute inflammatory damage to more diffuse neurodegeneration and widespread inflammation underlies the insidious accumulation of disability characteristic of progressive MS (Kuhlmann et al., 2023).

Current DMTs primarily target inflammation and have proven highly effective in reducing the frequency of relapses and new lesion formation. These include injectable agents such as interferon-beta and glatiramer acetate, oral therapies like fingolimod and dimethyl fumarate, and monoclonal antibodies, including natalizumab and anti-CD20 agents, such as ocrelizumab and ofatumumab (Hauser et al., 2017; Montalban et al., 2017). However, a substantial proportion of disability progression occurs independently of clinical relapses. While early initiation of high-efficacy DMTs is associated with better long-term outcomes and delayed progression, current DMTs can only partially protect against progression and the neurodegenerative components of MS (Hauser and Cree, 2020; McGinley et al., 2021). Ocrelizumab was the first approved therapy for PPMS, demonstrating modest efficacy in slowing disability progression (Montalban et al., 2017). Siponimod, a selective sphingosine-1-phosphate receptor modulator, showed benefit in active SPMS, though with relatively modest effects (Kappos et al., 2018). Thus, there remains a significant unmet need for therapies addressing progressive mechanisms of MS.

Emerging therapeutic strategies increasingly focus on CNS-penetrant agents and neuroprotective mechanisms (Piehl, 2021). Bruton's tyrosine kinase (BTK) inhibitors, which modulate both peripheral B cells and CNS-resident microglia, have shown promising early results in progressive MS cohorts (Dolgin, 2021). Additionally, remyelination-promoting agents, such as clemastine and opicinumab, are under investigation, though clinical benefits remain to be fully established (Cadavid et al., 2017; Green et al., 2017). These evolving therapies highlight the growing recognition of the multifaceted pathology of MS and the need for treatments targeting both acute inflammatory and chronic, neurodegenerative components. In line with this, this thesis also investigates the effects of the experimental drugs anakinra and tolebrutinib on chronic inflammation.

1.3 Conventional MRI of White Matter Lesions in Multiple Sclerosis

MRI is the most important paraclinical tool for diagnosing, monitoring, and investigating MS (Hemond and Bakshi, 2018). It plays a central role in all stages of clinical care, from the initial detection of lesions to assessing treatment response and guiding therapeutic decisions (Thompson et al., 2018; Wattjes et al., 2021). To ensure diagnostic accuracy and comparability across centres, standardized MRI protocols have been proposed by expert groups, including the European MRI in MS network (MAGNIMS), the Consortium of MS Centers (CMSC), and the North American Imaging in MS Cooperative (NAIMS) (Wattjes et al., 2021). MRI is recommended in all patients undergoing evaluation for MS. For routine clinical use, the recommended field strength is at least 1.5T, with 3T preferred when available due to its higher resolution and signal-to-noise ratio (SNR). However, 1.5T scanners remain widely used in clinical settings due to greater availability.

In this thesis, cMRI refers to a set of well-established and standardized protocols that are routinely used in clinical practice at 1.5T or 3T. The core sequences recommended for MS imaging include T2-weighted 3D fluid-attenuated inversion recovery (FLAIR), axial T2-weighted (T2w), and T1-weighted (T1w) imaging with and without gadolinium contrast (Wattjes et al., 2021). A key application of cMRI is the visualization of focal demyelinating lesions, which are most commonly located in the white matter (WM) of the CNS (Figure 1.2). FLAIR is particularly effective for detecting WML due to suppression

of cerebrospinal fluid signal. Axial proton density and T2w images provide additional information on lesion load and distribution and can be useful for problem-solving when there are artefacts on T2-FLAIR. Gadolinium-enhanced T1w images allow for identification of active inflammatory lesions, also known as enhancing lesions, which appear hyperintense due to transient BBB disruption. These lesions are indicative of ongoing inflammation and demyelination (Bruck et al., 1997; Frank et al., 1994; Lassmann, 2008; McFarland et al., 1992). Once the BBB resolves, usually within 2-8 weeks after lesion onset, lesions no longer show enhancement after gadolinium injection (Absinta et al., 2016a). Radiologically, these non-enhancing lesions are then classified as chronic lesions.

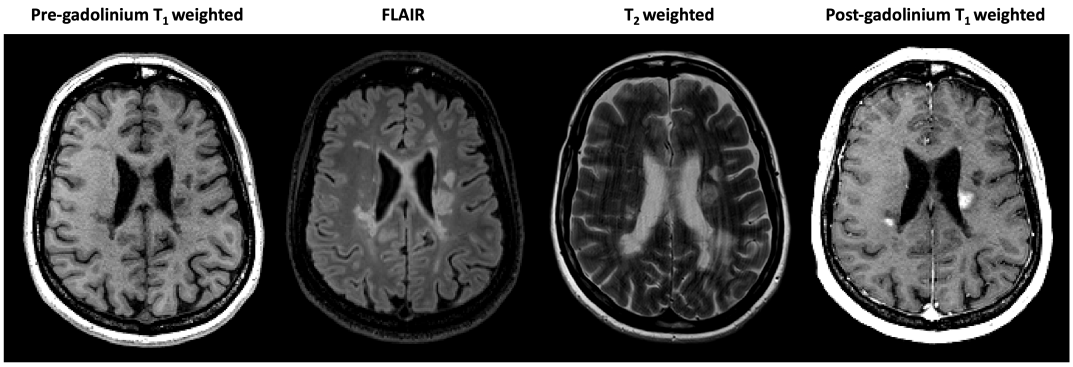


Figure 1.2: Recommended sequence contrasts for routine clinical scans for MS at 3T. From left to right, a pre-gadolinium T1-weighted, a 3D T2-weighted FLAIR, T2-weighted, and a post-gadolinium injection T1-weighted scan showing two enhancing lesions. Post-injection T1-weighted should be acquired at least 5 minutes after the bolus injection. This figure is reproduced from Donnay CA et al., Advanced MRI in Multiple Sclerosis, Elsevier, book chapter under review.

Lesion detection with cMRI plays a central role in the formal diagnosis of MS, particularly through its integration into the McDonald criteria, a set of guidelines used to diagnose MS (Thompson et al., 2018). Within this framework, the diagnosis of MS requires evidence of dissemination in time (DIT) and space (DIS), which can be demonstrated radiologically. DIT can be demonstrated either by the simultaneous presence of enhancing and non-enhancing lesions on a single scan or by identifying new lesions on follow-up imaging. MRI can demonstrate DIS by the presence of lesions in distinct locations within the brain (infratentorial, juxtacortical, cortical, periventricular) or the spinal cord.

Beyond diagnosis, cMRI is widely used to monitor disease activity and treatment response. Serial imaging is recommended before, during, and after initiation or switching of DMTs, with scan intervals adapted based on disease activity (Wattjes et al., 2021). The number and volume of visible lesions,

referred to as lesion count and lesion load, are associated with a more aggressive disease course and can help inform ongoing treatment decisions (Barkhof et al., 1999; Fisniku et al., 2008). Furthermore, the presence of enhancing lesions can serve as a meaningful predictor of treatment response (Wattjes et al., 2021) and is used as an endpoint in clinical trials.

1.4 Advanced MRI of White Matter Lesions in Multiple Sclerosis

While cMRI remains the clinical cornerstone for diagnosing and monitoring MS, it has important limitations that constrain both its interpretability and accessibility. First, cMRI offers limited insight into the biological processes driving disease progression. Sequences such as FLAIR and T1w imaging are highly sensitive to WML but cannot reliably differentiate between inflammation, demyelination, or axonal loss. Consequently, lesion appearance on cMRI provides only a partial view of ongoing tissue damage or repair, limiting its utility for understanding and tracking disease progression. Second, access to cMRI remains uneven across clinical settings (Solomon et al., 2023). Standard clinical scanners are resource-intensive, requiring specialized infrastructure, technical expertise, and ongoing operational costs. This limits their availability in low-resource environments and restricts frequent or widespread monitoring of neurological diseases. Together, these cMRI limitations have prompted the advanced MRI developments aimed at improving the biological specificity and expanding accessibility of MRI.

Advanced MRI encompasses both non-conventional sequences and imaging at non-standard field strengths. Advanced sequences that are sensitive to tissue microstructure, combined with UHF MRI's resolution and contrast, enable visualization of subtle pathological features. These capabilities allow the study of mechanisms relevant to progressive MS, addressing key gaps in understanding the disease. At the other end of the spectrum, ULF MRI systems offer a promising solution to the challenge of MRI's limited accessibility. By providing portable and cost-effective imaging, ULF MRI reduces reliance on resource-intensive high-field scanners and creates opportunities for broader, more frequent monitoring even in settings where standard MRI is not readily available.

Advanced MRI has contributed to many aspects of MS research and clinical care. The following section will focus on its application to lesion assessment, highlighting chronic lesion evaluation with 7T SSI and qT1 mapping, and more broadly exploring the potential of lesion assessment with ULF MRI.

1.4.1 7T MRI for Chronic White Matter Lesions Assessment

Advances in MRI have opened a window into the in-vivo characterization of chronic lesion heterogeneity, offering potential dynamic biomarkers of non-relapsing, progressive MS biology. This is particularly important for monitoring ongoing tissue injury that is not captured by cMRI's traditional markers of acute inflammation.

1.4.1.1 Paramagnetic Rim Lesions

After acute inflammation resolves, chronic lesions can evolve along multiple paths. Some lesions begin to remyelinate (Popescu et al., 2013), while other lesions remain chronically demyelinated and are further categorized into chronic inactive or chronic active lesions. Chronic inactive lesions are primarily characterized by "homeostatic" microglia, while chronic active lesions exhibit sustained, organized inflammation at their edges, marked by the accumulation of macrophages and activated microglia (Lassmann, 2018*b*; Luchetti et al., 2018). This persistent inflammatory activity reflects ongoing tissue injury and failed repair and has become increasingly recognized as a key contributor to disease progression (Kuhlmann et al., 2023).

While previously only identifiable pathologically, advancements in SSI techniques have facilitated the detection of chronic inflammation around the edges of chronic active lesions. SSI is sensitive to local variations in tissue magnetization caused by paramagnetic and diamagnetic substances. This enables the in vivo detection of iron-laden pro-inflammatory microglia and macrophages at the lesion edge, which appear as a paramagnetic rim (Absinta et al., 2018, 2016; Bagnato et al., 2024; Dal-Bianco et al., 2017; Hemond et al., 2022). Representative examples of PRL+ and PRL- are shown in Figure 1.3. Indeed, post-mortem studies have confirmed co-localization between the presence of the paramagnetic rim and the accumulation of iron at the edge of the lesion (Dal-Bianco et al., 2017; Gillen et al., 2021; Mehta et al., 2013; Yao et al., 2012).

Several SSI-based acquisition and post-processing strategies have been developed for PRL detection; however, there is currently no consensus within the field regarding the optimal imaging sequence or post-processing approach (Bagnato et al., 2024). At its simplest, SSI involves a gradient-echo T2*-weighted sequence (T2*-GRE), where the phase image distinguishes between paramagnetic and diamagnetic susceptibilities, as determined by local frequency shifts in the MRI signal. Both single and multi-echo T2*-GRE sequences have been used to identify PRLs. Additionally, T2*-GRE sequences remain heavily influenced by acquisition parameters such as echo time and the lesion's orientation relative to the magnetic field. Since susceptibility effects are anisotropic, rim visibility can vary depending on the lesion's alignment, and contrast may be reduced at lower field strengths. Vendor-agnostic post-processing pipelines employ a two-step processing approach involving phase unwrapping followed by high-pass filtering. This approach can greatly enhance PRL sensitivity, but it also introduces dipolar artefacts due to the geometry and orientation of nearby veins and lesions. These artefacts can result in misleading appearances, sometimes described as *pseudo-PRLs*, which complicate interpretation (Wisnieff et al., 2015).

Susceptibility-weighted imaging (SWI) improves upon T2*-GRE by combining magnitude and filtered phase data, thereby increasing paramagnetic rim and lesion core contrast and minimizing orientation dependence (Haller et al., 2021). Several vendor-supported post-processing variants of SWI exist, including susceptibility-weighted angiography (SWAN) and SWI-phase (SWIp), which apply proprietary algorithms to combine magnitude and phase data in varying ways.

With more sophisticated post-processing, quantitative susceptibility mapping (QSM) can quantify magnetic susceptibility variations and improve the differentiation between myelin and iron, which is sometimes ambiguous even after unwrapping and filtering the phase images (Wisnieff et al., 2015). QSM reconstructs maps of susceptibility values from multi-echo GRE phase data by de-convolving the phase image to remove background field contributions (Liu et al., 2015). Unlike qualitative SWI or filtered phase images, QSM enables the estimation of absolute susceptibility values, improving specificity for iron detection and allowing better differentiation between iron-related signal changes and those from other sources, such as myelin or de-oxygenated blood (Langkammer

et al., 2010, 2012). While QSM provides improved biological specificity, its implementation remains technically complex, requiring precise phase unwrapping, background field removal, and regularized inversion. These computational demands, along with variability in reconstruction pipelines, currently limit its routine use in clinical settings.

Currently, there is limited data available regarding the impact of different MRI sequences and post-processing techniques on PRL detection. However, 7T systems offer significantly improved susceptibility and T2* contrast, owing to the increased field-dependent susceptibility effects, and 7T is currently the gold standard imaging field strength for in-vivo PRL detection (Bagnato et al., 2024). Studies comparing field strengths have shown that PRLs are more reliably detected at 7T, with improved rim delineation, higher detection rates, and lower false positives and negatives relative to 3T (Absinta et al., 2018; La Rosa et al., 2022).

Nevertheless, guidelines for PRL identification have been published, outlining criteria across various acquisition types (Bagnato et al., 2024). According to these guidelines, PRLs are detected using SSI and are characterized by a T2w-hyperintense core that does not enhance with gadolinium, surrounded by a paramagnetic rim visible along at least two-thirds of the lesion perimeter. For a lesion to qualify as a PRL, part of the rim must be visible on at least two consecutive slices in 2D acquisitions or in two orthogonal planes in 3D acquisitions. In this thesis, these guidelines are applied to identify PRLs on GRE-T2 acquisitions for the longitudinal study of chronic lesions.

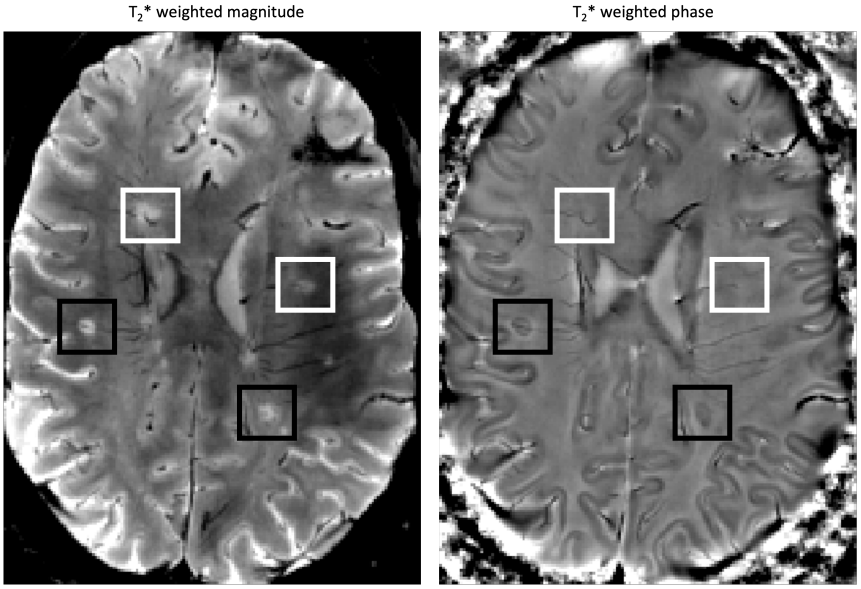


Figure 1.3: Chronic Lesions on T2*-GRE magnitude and phase sequence at 7T. Black boxes show lesions that have a paramagnetic rim lesion on phase. White boxes show lesions that do not have a paramagnetic rim lesion. This figure is reproduced from Donnay CA et al., Advanced MRI in Multiple Sclerosis, Elsevier, book chapter under review.

The clinical importance of PRLs has become increasingly apparent. For one, PRLs seem to be specific to MS (Maggi et al., 2020), aiding in the differential diagnosis of MS. In a multi-center study of 438 individuals, PRLs were found in 52% of MS cases compared to only 7% of individuals with other diagnoses (Maggi et al., 2020). This specificity makes PRLs useful in differentiating MS lesions from MS-mimicking conditions, especially when considered with other advanced MRI biomarkers like the central vein sign (Martire et al., 2022). However, not all MS patients have PRLs; a review of 29 studies by (Ng Kee Kwong et al., 2021) revealed that, on average, PRLs appear in 9.8% of all WML and about 40.6% of patients.

In addition to their diagnostic value, PRL+ appear to be more destructive than PRL-. The number of PRLs is linked to both the overall mean WML volume and the accrual of disability at a younger age (Absinta et al., 2019, 2016; Dal-Bianco et al., 2021a). This association with clinical disability is likely attributed to the destructive nature of PRLs. Cross-sectional analyses reveal that PRLs are larger and display heightened tissue damage, as indicated by myelin-sensitive quantitative MRI, and have poor remyelination (Kolb et al., 2021; Yao et al., 2018). Longitudinal imaging at 7T has shown that, compared to PRL-, PRL+ become increasingly T1 hypointense over time, suggesting that rim presence may mark an early failure of lesion repair and the onset

of irreversible tissue damage (Absinta et al., 2016). In later disease stages, these lesions are more likely to smoulder, increasing in volume and exerting deleterious effects on surrounding tissue (Absinta et al., 2019, 2016; Dal-Bianco et al., 2017).

Importantly, recent 7T studies have demonstrated that in a subset of lesions, the paramagnetic rim can shrink or even disappear over time (Absinta et al., 2021; Dal-Bianco et al., 2021), suggesting that rim dynamics could serve as a pharmacodynamic biomarker in clinical trials targeting chronic inflammation. Building on these findings, this thesis uses PRL resolution as an endpoint in the clinical trial analyses.

Despite their clinical relevance, there are several challenges associated with the detection and interpretation of PRLs. Variability in imaging parameters and methodological approaches across studies makes it difficult to standardize PRL assessment. A recent review summarised the reliability of per-subject PRL number as *good* on average (Reeves, Mohebbi, Zivadinov, Bergsland, Dwyer, Salman, Schweser and Jakimovski, 2023); however, intra- and inter-rater reliability remain variable. Additionally, expert visualization and identification remain the standard methods for quantification, yet this process is often laborious and time-consuming (Kolb et al., 2022; La Rosa et al., 2022). While automated detection algorithms for PRLs have been developed at 3T using T2*-GRE phase (Barquero et al., 2020; Lou et al., 2021) and QSM (Zhang et al., 2022), the main challenge for developing automated methods is the limited availability of datasets with manual annotations and the inter-rater variability of manual annotations (La Rosa et al., 2022).

1.4.1.2 Slowly Expanding Lesions

Another emerging marker of chronic lesion activity is the slowly expanding lesion (SEL), which reflects ongoing, smouldering inflammation within chronic MS lesions by capturing their slow, progressive volumetric expansion over time. SELs are identified through longitudinal post-processing techniques as lesions that locally expand over time without new gadolinium enhancement (Elliott, Belachew, Wolinsky, Hauser, Kappos, Barkhof, Bernasconi, Fecker, Model, Wei and Arnold, 2019; Elliott, Wolinsky, Hauser, Kappos, Barkhof, Bernasconi, Wei, Belachew and Arnold, 2019).

The typical method for identifying SELs involves non-linearly registering a baseline image to a follow-up scan, which enables the detection of local tis-

sue deformations over time. The typical method for identifying SELs involves non-linearly registering a baseline image to a follow-up scan, which enables the detection of local tissue deformations over time. The Jacobian matrix of the resulting deformation field at each voxel captures the local three-dimensional spatial derivatives of the displacement field. The determinant of these voxelwise Jacobian matrices is then computed, yielding a scalar value at each voxel that represents the local volumetric expansion or contraction at the follow-up timepoint relative to baseline. SEL candidates are identified as spatially contiguous regions within baseline T2 lesions that exhibit local volume expansion exceeding a predefined threshold. These candidate regions are subsequently refined by iteratively incorporating adjacent voxels that meet secondary expansion criteria and by applying a minimum size constraint to define final SELs (Elliott, Wolinsky, Hauser, Kappos, Barkhof, Bernasconi, Wei, Belachew and Arnold, 2019).

SELs have been shown to be an independent predictor of EDSS worsening (Preziosa et al., 2022). Additionally, some studies show that SELs are more commonly found in progressive types of MS, but these findings are inconsistent (Arnold et al., 2021).

Although SELs can be detected using cMRI sequences such as T1-w or T2-w imaging, the complex post-processing required for their identification currently limits their clinical feasibility. The reliability of SEL detection depends heavily on the quality of the MRI data and the accuracy of image processing steps; artefacts or errors in lesion delineation and registration can significantly influences volume change estimates. Moreover, SEL identification requires at least two timepoints for assessment, and incorporating additional scans beyond two timepoints becomes increasingly complex. Many of these challenges have yet to be thoroughly explored, emphasizing the need for more studies to comprehensively validate and refine the SEL detection method (Elliott, Wolinsky, Hauser, Kappos, Barkhof, Bernasconi, Wei, Belachew and Arnold, 2019).

While both PRLs and SELs are thought to reflect chronic active pathology and ongoing tissue destruction at the lesion border, they do not always detect the same lesions. In fact, a study examining 61 MS participants found that out of 1492 lesions, 80 PRLs were identified, and only 54% of PRLs co-localized with SELs (Calvi et al., 2023). Similarly, another study found a 39.5% overlap between PRLs and SELs (Elliott et al., 2023). Thus, while PRLs and SELs may partially overlap, they do not have equivalent sensitivity and specificity to

histopathological chronic active lesions. Among the two, PRLs currently have the most robust histopathological validation as an imaging biomarker of chronic active pathology (Bagnato et al., 2024). Accordingly, SELs are reviewed here for context, but only PRLs are directly considered in the experimental analyses of this thesis.

1.4.1.3 Quantitative T1 Mapping for Monitoring White Matter Lesions

MRI sequences sensitive to microstructural damage provide insight into the tissue environment of chronic lesions. An early example is the identification of *black holes* on T1w spin-echo images, where greater T1 hypointensity reflects irreversible axonal loss and correlates with disability progression (Bagnato et al., 2003; Kolb et al., 2021; Sahraian et al., 2010). Beyond this, advanced MRI sequences such as magnetization transfer imaging (MTI), myelin water imaging (MWI), quantitative T1 and T2 mapping, and diffusion-based models (e.g., DTI, NODDI) aim to capture aspects of myelin integrity and axonal structure (Basser, 1995; Caranova et al., 2023; Granziera et al., 2021). While not yet routine in clinical practice, these methods offer valuable insight into chronic lesion pathology. Among them, qT1 mapping currently shows the greatest technical maturity and clinical readiness for assessing myelin damage and repair (Granziera et al., 2021).

Quantitative T1 (qT1) relaxometry is used to measure the recovery of longitudinal magnetization of excited spins in a tissue. Unlike conventional T1w sequences, which provide images in arbitrary intensity units, qT1 mapping assigns each voxel an absolute T1 value. In the context of MS, qT1 mapping allows in vivo assessment of tissue integrity. Histopathological analysis of MS tissues has shown that qT1 has a strong and moderate correlation with myelin content and axonal density, respectively (Mottershead et al., 2003). Myelinated WM exhibits short qT1 times, while demyelination prolongs qT1 values (Brex et al., 2000; Schmierer et al., 2004). Axonal loss also contributes to qT1 increases, whereas iron accumulation, which shortens T1, may partially offset this effect (Jonkman et al., 2015; Stuber et al., 2014; van Waesberghe et al., 1999). While determining the relative contributions of myelin, axonal loss, and iron remains challenging (Brex et al., 2000; Seewann et al., 2009), demyelination appears to be the dominant factor contributing to initial T1 prolongation in MS (Kolb et al., 2021) and is used in this thesis's 7T studies of chronic WMLs.

Among the various approaches to T1 mapping, the Magnetization-Prepared Two Rapid Acquisition Gradient Echoes (MP2RAGE) sequence has gained prominence due to its efficiency and robustness. MP2RAGE acquires two inversion recovery images with different inversion times in rapid succession, generating both a synthetic bias-corrected T1-weighted image and a voxel-wise T1 map (Marques et al., 2010). The method provides uniform signal intensity that is relatively insensitive to proton density, T2* effects, and B1 field inhomogeneities, overcoming some of the limitations of conventional T1-weighted imaging (Kober et al., 2012). Despite these advantages, absolute T1 values may vary across different mapping techniques, highlighting the need for standardization (Granziera et al., 2021).

In MS research, qT1 mapping has demonstrated sensitivity in differentiating lesions from normal-appearing white matter (NAWM) and grey matter (Chen et al., 2023; Kober et al., 2012; Rahmanzadeh, Weigel, Lu, Melie-Garcia, Nguyen, Cagol, La Rosa, Barakovic, Lutti, Wang, Bach Cuadra, Radue, Gaetano, Kappos, Kuhle, Magon and Granziera, 2022). WMLs show longer qT1 values compared to surrounding tissues, with the highest values typically observed in the lesion core (Bonnier et al., 2019; Simioni et al., 2014). This spatial gradient mirrors histopathological findings of more severe tissue loss in the core than in the periphery. Prolonged qT1 in WMLs has been linked to higher disability in RRMS, while increased qT1 in cortical lesions correlates with cognitive decline (Chen et al., 2023; Harper et al., 2023; Simioni et al., 2014; Thaler et al., 2015). Chronic active lesions, known for their ongoing inflammatory activity and neurodegeneration, also demonstrate significantly prolonged qT1 values, supporting the role of qT1 as a marker of lesion severity (Kolb et al., 2021).

Additionally, the use of 7T qT1 mapping, particularly with the MP2RAGE sequence, has enabled more precise in-vivo tissue characterization of chronic lesions (Figure 1.4). The enhanced CNR, SNR, and the increased spatial resolution (0.5–0.7mm isotropic) at 7T improve signal separation, allowing for the detection of subtle changes in tissue composition that may reflect ongoing demyelination, partial remyelination, or axonal degeneration (Kolb et al., 2021).

Beyond cross-sectional characterization, longitudinal studies of qT1 offer promising insights into disease progression and treatment response. For example, a reduction in qT1 values in black holes over one year was associated

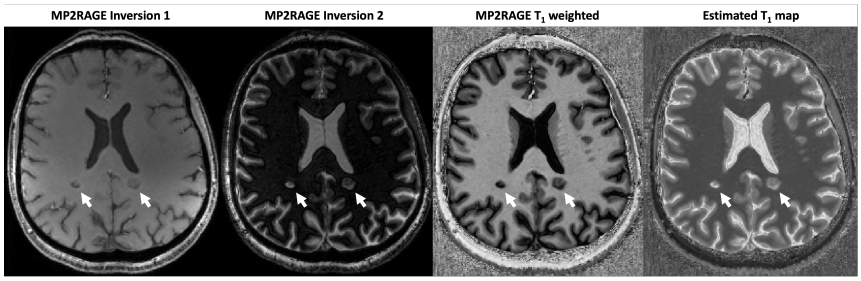


Figure 1.4: Example chronic lesions on 7T MP2RAGE contrasts at 0.5mm isotropic resolution. 7T MP2RAGE T1 maps show high separation of signal as can be seen in the example lesions on the estimated T1 map, where heterogenous qT1 likely indicate mixed amounts of demyelination within even a single lesion. This figure is reproduced from Donnay CA et al., Advanced MRI in Multiple Sclerosis, Elsevier, book chapter under review.

with clinical improvement and treatment response (Thaler et al., 2017). Similarly, (Rahmanzadeh, Galbusera, Lu, Bahn, Weigel, Barakovic, Franz, Nguyen, Spincemaille, Schiavi, Daducci, La Rosa, Absinta, Sati, Bach Cuadra, Radue, Leppert, Kuhle, Kappos, Bruck, Reich, Stadelmann, Wang and Granziera, 2022) reported that longitudinal increases in qT1 within WMLs, NAWM, and deep GM correlated with worsening disability as measured by EDSS. These findings suggest that qT1 mapping could serve as a valuable biomarker for clinical trials, particularly those targeting remyelination or neuroprotection. In this thesis, qT1 mapping is used to track PRL+ and PRL- lesion trajectories in both natural history and clinical trial 7T datasets.

1.4.2 Ultra-Low Field MRI for White Matter Lesion Assessment

ULF MRI, operating at a field strength below 0.1T, offers an accessible alternative to cMRI systems. While cMRI and UHF MRI provide increased spatial and signal resolution, they are expensive (often approximated at $1 M per tesla) and require specialized infrastructure such as shielding, helium, and high power (Arnold et al., 2023). These requirements limit availability, especially in rural or resource-limited settings and for patients with mobility impairments. In contrast, ULF scanners can, in principle, be manufactured at substantially lower cost due to simpler hardware. ULF MRI scanners typically use small permanent magnets or electromagnets, require no cryogens, and run on standard electrical outlets with modest cooling (Sarracanie et al., 2015). Some systems weigh less than 100kg and are mobile enough to be used at the bedside or in vans for point-of-care imaging (Deoni, Medeiros, Deoni, Burton, Beauchemin, D'Sa, Boskamp, By, McNulty, Mileski, Welch and Huentelman, 2022). In MS,

ULF MRI could support more frequent imaging, imaging in under-served communities, and for patients who cannot travel to a radiology suite.

Currently, several ULF brain scanners exist or are under development (Altaf et al., 2024; Broche et al., 2019; Cooley et al., 2021; Liu, Leong, Zhao, Xiao, Mak, Tsang, Lau, Leung and Wu, 2021; O'Reilly et al., 2019; Sarracanie et al., 2015). The most established and used in this thesis is the Hyperfine Swoop system (www.hyperfine.io), a 64mT scanner cleared by the U.S. Food and Drug Administration in 2020 for point-of-care brain imaging. The Hyperfine Swoop system weighs approximately 630 kg and is used in hospital settings. The company offers sequences standardized for the scanner's low field, including T1w, T2w, FLAIR, and basic diffusion sequences for brain imaging, which are acquired at a coarse resolution (1.5–1.8mm in-plane, 5mm slice thickness). The sequences are offered in axial, coronal, and sagittal planes and can be acquired in $\sim$10 minutes. The scanner software includes automated reconstruction and basic visualization; however, these are constrained by a closed design. Users receive *DICOM* images but cannot readily export raw k-space for custom reconstructions. Additionally, the ULF sequences are locked, and users cannot modify gradient strengths or pulse shapes. Although the Hyperfine Swoop system features built-in electromagnetic interference cancellation, it can still be affected by environmental interference (Islam et al., 2023).

Increasingly, studies have focused on evaluating the clinical utility of portable ULF MRI systems in MS, examining both their integration into clinical workflows and their ability to detect relevant imaging biomarkers. Clinically, studies have shown that incorporating portable ULF MRI can reduce the time from symptom onset to initial imaging (Lim et al., 2024) and has been well tolerated by people with MS compared to HF MRI, due to its open design and quieter operation (Ham et al., 2024). In terms of detecting relevant imaging biomarkers, ULF MRI has demonstrated the ability to reliably detect demyelinating lesions larger than 4mm in diameter (Arnold et al., 2022). A recent study involving 55 MS cases found that neuroradiologists, blinded to clinical and high-field MRI data, were able to detect WML with high sensitivity (Okar et al., 2025). While this was particularly true for periventricular lesions, higher false-positive rates were found in juxtacortical areas. Notably, self-training using a curated case set significantly improved detection performance, enabling raters to correctly identify DIS based on periventricular and juxtacortical lesions alone. While

these findings are promising, further studies are needed to validate the reliability of ULF MRI for WML monitoring and to define its clinical role in MS care.

1.5 Challenges in Studying White Matter Lesions with Advanced MRI

MRI biomarkers play a crucial role in quantifying disease burden, monitoring disease progression, and evaluating treatment response. The quantification of MRI WML biomarkers is crucial in clinical trials and research in MS. Reliable image processing methods are therefore fundamental to ensure the validity and reproducibility of results. Despite the growing reliance on automated methods, there is limited research examining how different analysis pipelines influence WML analyses. The variability introduced by specific tools and processing steps is further compounded when using non-conventional field strengths, such as UHF or ULF MRI, where deviations in contrast, resolution, and image quality from conventional MRI can cause standard neuroimaging tools to fail or produce unreliable results. This section provides an overview of WML quantification in MS, introducing specific challenges related to UHF and ULF MRI. More detailed literature reviews are included in the relevant experimental chapters.

In this thesis, I build on these studies by evaluating the potential of ULF MRI, using the Hyperfine Swoop system, for WML detection and longitudinal monitoring in MS.

1.5.1 Quantification of White Matter Lesions

Two common measures of WMLs are lesion count and lesion burden, both of which have been associated with long-term disease outcomes, including physical and cognitive disability progression (Brex et al., 2002; Goodin et al., 2012; Ouellette et al., 2018; Uher et al., 2017). Lesion count refers to the number of distinct lesions visible on MRI, while lesion burden (also referred to as lesion load) refers to the total volume of lesional tissue. These WML measures can be quantified using a range of approaches, from visual radiological assessment to semi-automated and fully automated methods. Radiological assessment of WML is routinely reported in clinical practice (Alessandrino et al., 2018; Wattjes et al., 2021), with guidelines recommending the reporting of either the

total number of T2-hyperintense lesions or lesion burden using bracketed volume ranges (Wattjes et al., 2021). The number of WML is typically counted manually by experienced reviewers. However, this process is time-consuming and particularly challenging in patients with high lesion burdens or confluent lesions that are difficult to distinguish individually (Zivadinov et al., 2008). Additionally, the reproducibility of expert lesion counting is only moderate and depends on lesion location, with one study showing higher inter-rater agreement in supratentorial regions and in cases with low total lesion load (Bozsik et al., 2022). However, inter-rater agreement improves when predefined lesion count categories are used instead of exact lesion numbers (Barkhof et al., 1999).

In addition to lesion counting, lesion burden can be quantified through segmentation of lesional tissue, where individual voxels in an MRI scan are labeled as WML. Manual segmentation remains the traditional and most accurate method for quantifying lesion volume. However, similar to radiological assessments of WML burden and WML count it is highly time-consuming, labour-intensive, and prone to intra- and inter-rater variability (De Rosa et al., 2024; Garcia-Lorenzo et al., 2013).

Lesion segmentation masks can also be useful to automate the processes of counting lesions. At its most basic, lesions can be counted as the number of distinct spatial clusters in the WML mask. However, this pipeline may fail in the presence of diffusely abnormal WM (DAWM) and confluent lesions, both found in MS. DAWM are non-focal areas with MRI abnormality where the signal intensity lies between NAWM and classic WML (Cairns et al., 2022). Confluent lesions are lesions that arise separately but occur close to one another, resulting in a larger interconnected area of demyelination (Zivadinov et al., 2008). Both complicate the separation of individual lesions because they obscure defined boundaries. Automated efforts have aimed to address these challenges by using statistical analysis techniques to count lesions (Dworkin et al., 2018) or by creating distinct segmentation classes for DAWM and WML lesions (Maranzano et al., 2020). However, consistent quantification remains challenging due to the variability in lesion appearance and size, as well as the presence of confounding factors such as noise and artefacts in MRI acquisitions.

Automating lesion segmentation can improve the efficiency of lesion count and lesion burden. Automated segmentation algorithms typically generate a segmentation mask, delineating the total lesion area as a distinct class from the

rest of the MRI. A wide range of lesion segmentation tools have been developed using different methodological approaches. These include semi-automated techniques such as thresholding, fuzzy connectedness, and seed-based region growing, which assist in the delineation process but still require considerable manual oversight and often struggle with large, complex datasets (Ma et al., 2022). Fully automated algorithms, including both supervised and unsupervised machine learning models, have gained popularity due to their scalability and performance. Supervised methods require annotated training data and can generalize poorly to unseen scanners or acquisition protocols, whereas unsupervised methods avoid this dependency but may be less accurate overall. There have been multiple reviews of WML segmentation algorithms in MS, which suggest that supervised methods tend to outperform other approaches despite their limitations (Danelakis et al., 2018; Zeng et al., 2020). However, comparing segmentation techniques across studies remains difficult because few public datasets with ground truth are available, and many algorithms are tailored to the specific datasets on which they were developed. Convolutional neural networks (CNNs), particularly U-Net architectures, have shown promise in improving segmentation accuracy and are now widely used in research settings (Bonacchi et al., 2022). However, challenges such as high lesion variability, scanner heterogeneity, and the tendency for automated methods to underperform on novel data continue to limit their clinical utility. Despite ongoing advances, no automated lesion segmentation tool has yet achieved sufficient robustness and reliability for routine clinical use, and expert review of lesion segmentation is still advised (Dworkin et al., 2018; McKinley et al., 2020; Zeng et al., 2020). These challenges are further exacerbated at non-standard field strengths. For this reason, this thesis addresses automated segmentation at both UHF and ULF, leveraging U-Net architectures and transfer learning approaches.

1.5.2 Tracking Longitudinal Changes in White Matter Lesions

Quantifying longitudinal tissue volume changes is essential to understanding disease progression and therapy response. In MS, radiologists routinely conduct serial MRI scans to track lesion activity, visually assessing the presence of new, enlarging, shrinking and stable lesions relative to previous time points. Even within clinical practice, the integration of some image processing steps has

emerged as a valuable tool for evaluating longitudinal WML changes, although they are more commonly employed in research and clinical trials. Automated WML longitudinal evaluation relies heavily on accurate image processing steps such as segmentation and registration.

Registration is the process of spatially mapping two or more images into a single coordinate system (Im et al., 2024). It requires selecting parameters that define the constraints on the transformations allowed for a particular registration scenario. For instance, rigid-body transformation is a linear transformation that only permits images to be translated and rotated along three axes. Rigid-body transformation is ideal for intra-subject registration because it does not allow the size or shape of the brain and other anatomical structures to change, which is generally true for within-subject MRIs (although geometrical distortions in some sequences relative to others may require more advanced transformations). In contrast, non-linear registration accommodates spatially varying deformations, enabling local bending and stretching. This offers the ability to capture complex local changes, a feature distinct from linear registration where a uniform transformation is applied to the entire image.

Several approaches have been developed to support longitudinal WML analysis in MS (Im et al., 2024). Deformation-based techniques, such as those used to identify SELs, leverage non-linear registration to capture local tissue changes (Elliott, Wolinsky, Hauser, Kappos, Barkhof, Bernasconi, Wei, Belachew and Arnold, 2019). Subtraction imaging, another common strategy, subtracts registered images voxel-wise to identify lesion growth, resolution, or shrinkage across timepoints (Horsfield et al., 2016; Moraal et al., 2009, 2010; Patel et al., 2017; Tan, van Schijndel, Fazekas, Filippi, Freitag, Miller, Yousry, Pouwels, Ader and Barkhof, 2002; Tan, van Schijndel, Pouwels, Ader and Barkhof, 2002). Semi-manual methods often begin with lesion segmentation at baseline, followed by linear registration of follow-up scans and manual refinement of lesion masks to capture new or evolving lesions (Beck et al., 2023; Klistorner et al., 2018).

The choice of longitudinal WML analysis strategy depends on the specific study goals and available resources. Automated pipelines are highly relevant for large-scale longitudinal studies and clinical trials, such as the 7T cohorts examined in this thesis, where scalability and reproducibility are essential. Subtraction imaging approaches are particularly valuable for radiological assessment and remain a focus of the ULF longitudinal analyses in this work. However, a common limitation across current longitudinal approaches is that they are

generally developed and or use tools developed for cMRI, making it difficult to perform robust WML analysis at non-standard field strengths like UHF or ULF.

1.5.2.1 Challenges of Lesion Analysis at 7T

At UHF, sub-millimetre resolution poses challenges for image processing, as most tools are designed for standard $1mm^3$ voxel sizes. Higher spatial resolution may reveal anatomical structures such as the dura mater and cerebral vasculature, which are not clearly visible in cMRI sequences (Bazin et al., 2014) and can interfere with WML segmentation and registration tasks.

Segmentation tools developed for cMRI may also underperform at UHF due to altered tissue relaxation properties at UHF (Vachha and Huang, 2021). At 7T, T1 values are prolonged, while T2 and T2* are shortened, leading to changes in image contrast and shifting intensity distributions away from the assumptions of many standard segmentation models. While some efforts have been made to adapt segmentation pipelines for 7T, their performance remains inconsistent across studies and imaging protocols (Ashburner and Friston, 2005; Bazin and Pham, 2008; Fujimoto et al., 2014; Goebel, 2012; La Rosa et al., 2020).

1.5.2.2 Challenges of Lesion Analysis at Ultra-Low Field

Similar to 7T, image processing at ULF presents distinct challenges, particularly when using commonly available neuroimaging software packages (Deoni et al., 2021; Deoni, O'Muircheartaigh, Ljungberg, Huentelman and Williams, 2022; Iglesias et al., 2021). At 0.064T, tissue magnetization and relaxation properties differ substantially from those at clinical field strengths. SNR decreases roughly linearly with field strength, requiring coarser spatial resolution to maintain acceptable image quality within reasonable scan times (Arnold et al., 2023). ULF MRI scans from the Hyperfine Swoop system exhibit elongation and stretching artefacts in the lower portion of the head and upper neck, consistent with spatial encoding errors arising from gradient nonlinearity outside the scanner's optimal field-of-view. This can introduce complications for image segmentation and registration. While the available sequences on systems like the Hyperfine Swoop (e.g., T1w, T2w, and FLAIR) reflect standard clinical sequences, the reduced tissue contrast and altered signal properties at

ULF make it difficult to distinguish lesions from NAWM (Arnold et al., 2023, 2022).

To address the resolution, SNR and CNR limitations of ULF, super-resolution (SR) and image synthesis methods have become areas of active research (Arnold et al., 2023). These approaches aim to enhance ULF images into higher-quality representations that are compatible with standard clinical MRI analysis pipelines. In principle, this enables the application of conventional neuroimaging tools (like tissue segmentation or registration).

Basic resolution enhancement can be achieved by registering and averaging orthogonal images or repeated acquisitions (Arnold et al., 2022; Deoni, O'Muircheartaigh, Ljungberg, Huentelman and Williams, 2022), but this requires long scan times and accurate registration, which may be challenging with ULF image quality. More sophisticated methods include CNNs. SynthSR is a CNN that generates a synthetic $1mm$ isotropic Tw image from input scans of any contrast, resolution, or orientation, including those from ULF systems (Iglesias et al., 2023a, 2021). These outputs can be processed using standard segmentation tools. SynthSeg, a complementary tool, segments brain structures without retraining and performs better when used after SynthSR (Billot et al., 2023). An extension, WMH-SynthSeg, enables segmentation of white matter hyperintensities (Iglesias et al., 2023b). Related image enhancement algorithms include LoHiResGAN a generative adversarial networks (GAN) trained on paired 64mT and 3T data that maps ULF images to synthetic high-field images using perceptual loss functions to preserve structural details (Islam et al., 2023), and the use of conventional and cyclic GANs on unpaired 0.064T and 1.5T data (Yoo et al., 2021).

There is a clear unmet need for dedicated tools to enable accurate WML analysis at UHF and ULF MRI field strengths. Developing and validating pipelines for these non-standard modalities expands the methodological framework for studying lesion evolution in MS. This thesis leverages longitudinal datasets acquired at the National Institutes of Neurological Disorders and Stroke (NINDS), where the Translational Neuroradiology Section (TNS) routinely collects advanced 7T and 0.064T MRI. These datasets include an MP2RAGE acquisition (providing a T1w image and a quantitative T1 map) and T2*-weighted imaging at 7T, as well as paired same-day 3T and ULF T1w and FLAIR scans. Together, these data offer a unique opportunity to study PRL

evolution and evaluate investigational treatments at 7T, while assessing the potential of ULF MRI for detecting WMLs relative to clinical standards.

The following section outlines the specific objectives of this thesis.

1.6 Summary of Research Objectives

Objective 1. Develop and evaluate automated 7T analysis methods for longitudinal analysis of qT1 and volume in chronic WML in MS.

1. Train and evaluate an automatic whole-brain and lesion segmentation model using 7T MP2RAGE images.
2. Assess the impact of different longitudinal analysis pipelines on the reliability of lesion qT1 and volume measurements.
3. Develop a visualization application to support qualitative, radiological assessment of paramagnetic rim lesions across time points and modalities.

Objective 2. Investigate longitudinal changes in qT1, volume, and paramagnetic rim characteristics in chronic MS lesions.

1. Quantify longitudinal changes in qT1 and lesion volume in a 7T natural history MS dataset.
2. Evaluate whether investigational treatments anakinra and tolebrutinib affect lesion volume, qT1, or paramagnetic rim.

Objective 3. Adapt and evaluate lesion analysis tools for ULF MRI.

1. Modify and assess segmentation algorithms for detecting white matter lesions on ULF MRI, relative to 3T.
2. Develop a subtraction-based longitudinal pipeline for ULF MRI to assess lesion evolution and compare radiological findings to 3T.
3. Improve ULF spatial resolution by combining spatial frequency information from orthogonal ULF acquisitions.

Objective 4 (Exploratory). Explore whether lesion tools developed for ULF MRI can capture clinically relevant changes observed on 3T MRI.

Chapter 2

Development of 7T MRI Analysis Tools for Longitudinal Tracking of Multiple Sclerosis Lesions

Contents

2.1 Development of 7T Whole Brain and Lesion Segmentation from a Single MRI Modality 30

2.2 Impact of Longitudinal Analysis Pipelines Quantitative T1 and Volume 47

2.3 Development of a Visualization Application to Streamline Multi-modal Longitudinal Lesion Assessment . 64

This chapter aims to develop tools to analyse longitudinal changes of quantitative T1 times (qT1) in chronic white matter lesions (WML) with or without a visible paramagnetic rim. This chapter is divided into three sections, each focusing on a specific methodological aspect and the corresponding tool required to study these changes. The results from this chapter will inform the methods in Chapter 3.4.3.

Section 2.1 introduces PLAn 7T (Pseudo-label Assisted nnU-Net), a whole-brain and lesion segmentation algorithm developed for 7T Magnetization-Prepared 2 Rapid Acquisition Gradient Echoes (MP2RAGE) MRI scans. In this section, I introduce and evaluate PLAn's performance compared to other segmentation algorithms, focusing on its ability to segment WMLs effectively. This tool automates skull-stripping, an important preliminary step for other analysis pipelines, and the identification of voxels corresponding to grey matter

(GM), white matter (WM), cerebral spinal fluid (CSF), and WMLs in each scan. The resulting WML segmentation serves as the foundation for subsequent steps, including tracking lesion changes over time and identifying lesions with a paramagnetic rim. The initial development and evaluation of 7T PLAn was carried out in collaboration with a colleague (H.D.), who contributed to the implementation of transfer learning strategies and optimization of learning rates. Dataset curation and evaluation for Cohort 1 (Figures 2.2-2.4) were performed jointly. Some of the results from this section, including both text and figures, were presented at the 2024 ISMRM conference (Donnay et al., 2024) and published in a co-first author paper with H.D. (Donnay et al., 2023).

While PLAn 7T offers a cross-sectional tool for segmenting WML at 7T, robust longitudinal analysis pipelines are necessary to track the evolution of qT1 and lesion volumes. Section 2.2 evaluates different longitudinal analysis pipelines using a test-retest dataset. The goal is not only to compare the performance of these pipelines but also to quantify the uncertainty in individual lesion qT1 and volume measures. Understanding these uncertainties is crucial for selecting the most appropriate analysis pipeline and for interpreting the reliability of our measurements when applying the pipeline to clinical datasets. The results in this section have been presented at the 2022 ACTRIMS and ECTRIMS conferences (Donnay et al, 2022).

Finally, in Section 2.3, I develop a visualization application designed to streamline the radiological rating of lesions, particularly for distinguishing between paramagnetic rim-positive (PRL+) and rim-negative (PRL-) lesions. While automated PRL classification algorithms exist (Barquero et al., 2020), due to the subtle nature of the paramagnetic rim, expert radiological evaluation remains the preferred approach for accurate PRL detection.

2.1 Development of 7T Whole Brain and Lesion Segmentation from a Single MRI Modality

2.1.1 Introduction

As outlined in Chapter 1.6, accurate volumetric segmentation of cortical and subcortical structures plays a pivotal role in tracking disease progression and is fundamental to the success of image processing pipelines. MRI is highly

sensitive in detecting focal MS lesions in the brain. Lesion segmentation is essential for tracking changes in lesion volume and extracting imaging biomarkers within lesioned areas. Manual lesion segmentation is accepted as the gold standard but is time-consuming and suffers from poor intra- and inter-expert reliability (Despotovic et al., 2015). While many automated or semi-automated segmentation algorithms are available, most are tailored for data acquired at conventional field strengths like 3T or 1.5T and often rely on atlases that do not account for MS-specific pathologies such as lesions, atrophy, or enlarged ventricles. Tools like *FreeSurfer*, which use probabilistic atlas-based methods, generally yield favourable results on scans of normal-appearing brains but may fall short when applied to anatomical variations seen in diseases such as MS (Selvaganesan et al., 2019).

Multi-contrast segmentation methods often leverage the value of T2-weighted fluid-attenuated inversion recovery (FLAIR) images for lesion detection. FLAIR sequences are useful for differentiating lesions from CSF, a key step in many lesion segmentation algorithms. One example is the Classification using DErivative-based Features (C-DEF) method (Selvaganesan et al., 2019). C-DEF is trained on skull-stripped, co-registered, and intensity-normalized 3T MRI modalities and then employs a logistic regression model based on image features extracted from MRI intensity derivatives to provide GM, WM, CSF, and WML segmentations. Importantly, C-DEF does not perform skull-stripping as part of its algorithm; instead, it relies on AFNI's *3d-skull-stripping* tool as a preprocessing step before applying the segmentation algorithm. This atlas-free approach has demonstrated superior performance to FreeSurfer in cases of high MS lesion load or brain atrophy (Selvaganesan et al., 2019).

While C-DEF has demonstrated strong performance at 3T, it has not been validated on ultra-high field (UHF) data such as 7T. Several factors may limit C-DEF's effectiveness when applied to 7T MRI data. Multi-modal segmentation methods like C-DEF rely on accurate registration across modalities, but this becomes more challenging at UHF due to increased spatial deformations and the potential for degraded co-registration quality (Spini et al., 2020). Additionally, FLAIR images, which are typically useful for lesion detection at conventional field strengths, may have reduced utility at 7T. This is due to factors such as diminished lesion conspicuity and stronger bias fields arising from multiple RF pulses, both of which can hinder accurate segmentation (Spini et al., 2020; Zwanenburg et al., 2010).

Deep learning (DL) segmentation models are a compelling tool for medical image segmentation (Hesamian et al., 2019). With the increasing computational power available, convolutional neural networks (CNNs) and their variants, such as U-Nets (Ronneberger et al., 2015), have become among the most effective methods for brain MRI segmentation tasks and have gained significant popularity in MS lesion segmentation (Svanera et al., 2021). These methods rely on layers of automatically optimized filters and non-linear activations to learn complex image features from annotated training datasets. The trained models are then applied to predict tissue segmentations on unseen data. Additionally, CNNs can be trained on multi-modal images, allowing researchers to tailor models to specific datasets.

While some studies have applied CNNs to segment 7T MRI data (Svanera et al., 2021), only a few have explored their use for MS-specific segmentation tasks at 7T, such as cortical MS lesion segmentation (La Rosa et al., 2021). A major draw-back of neural networks is that they often require tedious manual configuration and hyper-parameter tuning with adjustments needed for each new dataset. One approach that addresses this challenge is the "no new U-Net" (Isensee et al., 2021), an open-source 3D biomedical image segmentation tool designed to automate many of these steps. Based on the standard U-Net architecture (Ronneberger et al., 2015), the nnU-Net automatically configures the preprocessing, network architecture, training, and post-processing. It extracts dataset properties, like image size, voxel information, and class proportions, to automatically tune hyper-parameters and guide network construction and data manipulation. Following 5-fold cross-validation, nnU-Net selects the best-performing model from three configurations: a 2D U-Net, a full-resolution 3D U-Net, and a 3D U-Net cascade. This framework is effective across heterogeneous datasets, making nnU-Net a promising approach for segmentation tasks at 7T.

The nnU-Net, like other CNNs, still requires a training dataset, which is typically created through manual segmentations. Creating training labels is particularly challenging at ultra-high field strengths, where MRI images have small voxel sizes (hence many slices), making the creation of training labels tedious and time-consuming. Transfer learning could address this challenge, as it enables domain adaption where knowledge can be transferred across related tasks to improve performance of a new task. Transfer learning can be achieved using a pre-training and a fine-tuning step. First, the model is pre-trained on a

larger dataset that is related to the target task at hand, allowing it to pick up general features that are useful in the target task. Then, during fine-tuning, the network weights are adjusted using a smaller task-specific training dataset. As previous research suggests that transfer learning is effective when training data in the target domain is limited (Valverde et al., 2021), this approach could reduce the need for large amounts of annotated data in the target domain. Since 3T segmentation algorithms are easier to access and more readily available compared to 7T segmentation tools, one promising strategy is to pre-train a model using 3T data and then fine-tune it with 7T data. To our knowledge, this approach of transfer learning across MRI field strengths has not yet been tested for MS lesion segmentation tasks but could provide a solution to the data scarcity issue of 7T manual segmentations.

In this section, the development and evaluation of a 3T-to-7T transfer learning model, PLAn 7T, are described. PLAn is a 7T, whole-brain tissue (GM, WM, CSF) and WML segmentation model which uses the nnU-Net framework and a single MRI modality. Specifically, the tool was designed for use with the 7T MP2RAGE sequence, which is routinely acquired at the translational neuroradiology section (TNS) at the national institute of neurological disorders and stroke (NINDS). This approach aims to leverage the availability of 3T data, and their corresponding segmentations generated by existing automated segmentation algorithms (C-DEF trained on 3T data, referred to as C-DEF 3T). The performance of PLAn for the segmentation of whole brain, tissues and WML is first compared against three approaches: *FreeSurfer*, a C-DEF model trained on 7T data (C-DEF 7T), and a nnU-Net model trained on 7T data (nnU-Net 7T). Additional evaluations are performed comparing nnU-Net 7T and PLAn 7T.

2.1.2 Methods

2.1.2.1 Datasets for Training and Evaluation

Data used in this section were obtained from 57 participants enrolled in a Natural History protocol (Appendix A.1). Figure 2.1 provides an overview of how the data were grouped for training and evaluations. These dataset cohorts were used at different stages of algorithm development and evaluation, as described below. The specific dataset and cohort details described in Appendix A.2.

Paired 7T and 3T Dataset. In 40 participants, 7T and 3T scans were acquired within 9 months of each other (median interval= 36 days; range= 1−196 days). Participants did not exhibit clinical disease progression or develop new lesions between scans, allowing the 3T scans to serve as a reliable reference for evaluating 7T lesion segmentations (evaluation details in Section 2.1.2.3). These 40 participants were divided into Paired Cohort 1 ($n = 25$) and Paired Cohort 2 (n = 15), as shown in Figure 2.1A.

Five participants from Paired Cohort 1 were used as a training dataset. Segmentation labels for this training dataset ($n = 5$) were created to train C-DEF 3T, C-DEF 7T, nnU-Net 7T and fine-tune PLAn 7T. 3T training labels for C-DEF 3T were semi-automatically created as previously described (Cercignani and Wheeler-Kingshott, 2019; Selvaganesan et al., 2019). To create training labels for the 7T training group, FreeSurfer was run on the 7T MP2RAGE T1w image to generate a starting point for GM, CSF, and WM segmentations which were then manually edited for errors. FreeSurfer was chosen as a training segmentation starting point because it can be used "out of the box" (no re-training required for a new dataset). Having a starting point significantly reduces the time required to produce training labels across all tissue classes, especially given the high-resolution data. Nonetheless, substantial edits were required, including correction of misclassified tissue boundaries, missing WM detail in the cerebellum, and major skull-stripping errors, such as excessive removal of non-brain tissue or non-inclusion of sizeable brain regions. Finally, a manually drawn lesion mask and skull stripping mask (C.D., H.D.) was added. The final brain segmentation masks were validated by a trained and experienced neurologist (E.S.B. or M.I.G.).

For the remaining 20 participants in Paired Cohort 1 and all 15 participants in Paired Cohort 2, reference segmentation labels, "pseudo-labels" were created, using their 3T data. These pseudo-labels were created by segmenting 3T scans using C-DEF 3T and up-sampling the output segmentations to $0.7mm$ isotropic. The C-DEF 3T method was chosen as it has been previously shown to be accurate for 3T segmentation (Dieckhaus et al., 2022; Selvaganesan et al., 2019). In Paired Cohort 1, the pseudo-labels and corresponding up-sampled 3T images were used as an additional training dataset for PLAn-7T (see 2.1.2.2 for details). Additionally, in both Paired Cohort 1 and Paired Cohort 2, the pseudo-labels were used as reference segmentations to evaluate whole-brain and lesion segmentations from 7T algorithms, run on their paired 7T scans. This

was done to allow for evaluation of 7T algorithms using a reliable reference point without large-scale manual annotation.

7T Dataset. A total of 17 participants were only scanned at 7T (Figure 2.1A). Eight participants were used as a validation cohort for WML segmentation (7T Validation Cohort). For this cohort, manual WML segmentations were annotated on the MP2RAGE T1w image using ITK-snap and subsequently validated by a trained neurologist (E.S.B. or M.I.G.).

For the other 9 participants, test-retest data for 7T MP2RAGE scans were collected to assess the reproducibility of PLAn 7T and nnU-Net 7T (7T Test-Retest Cohort). One participant was excluded from the analysis due to motion artifacts detected in the second scan.

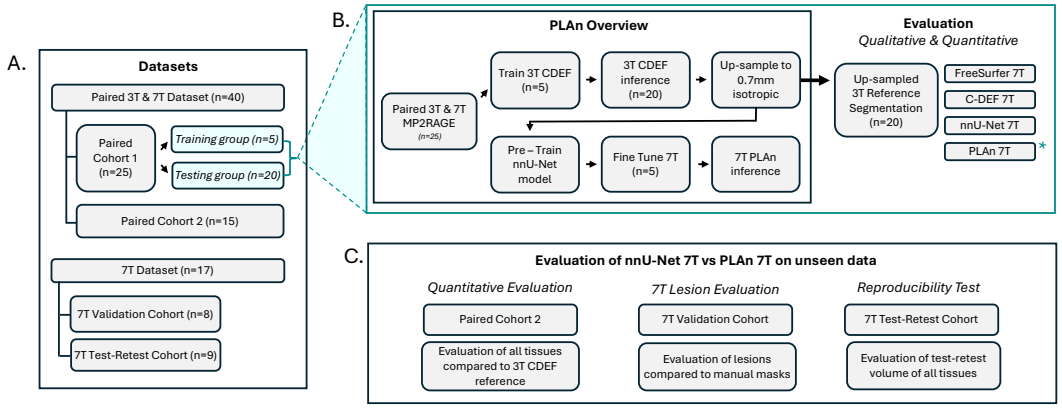


Figure 2.1: Overview of Datasets, PLAn-7T and Evaluation. A) Paired 3T and 7T, and 7T datasets used for training and evaluation. B) Diagram of the PLAn 7T segmentation algorithm and its initial evaluation of whole brain, tissues and lesions segmentation against other 7T segmentation algorithms. C) Overview of evaluation on unseen data performed between nnU-Net and PLAn 7T.
**Indicates biased evaluation.*

2.1.2.2 Overview of 7T Segmentation Approaches

An overview of all evaluations of the 7T segmentations methods is shown in Figure 2.1B and Figure 2.1C. FreeSurfer 7T, C-DEF 7T, nnU-Net 7T, and PLAn 7T were first qualitatively and quantitatively evaluated using the testing dataset ($n = 20$) from Paired Cohort 1 (Figure 2.1B). Since these same 20 participants were used to generate 3T pseudo-labels for pre-training PLAn 7T, this initial evaluation was slightly biased for PLAn 7T. To address this, the quantitative analysis was extended to include the 15 participants from Paired Cohort 2 to compare PLAn 7T with nnU-Net 7T (Figure 2.1C). Finally, using

the 7T Cohort, lesion segmentation and reproducibility of segmentations for both PLAn 7T and nnU-Net 7T were further assessed (Figure 2.1C).

For a qualitative analysis of Paired Cohort 1, segmentation outputs were randomly scrambled across methods, and the method used to generate each segmentation was concealed. Two experienced neurologists (M.I.G., C.T.) independently rated the segmentation quality for each tissue class and skull stripping on a scale from 1 to 5, with 5 representing the best segmentation (Table 2.1). Each method's mean scores and 95% confidence intervals were calculated and compared across the 20 participants.

Table 2.1: Qualitative Rating Definitions for Segmentation Evaluations.

Rating	**WML**	**Brain Tissue (WM/GM/CSF/skull-stripping)**
1	0-25% segmented, 100-75% not segmented or segmented as other structure	Most tissue misclassified
2	25-50% segmented, 75-50% not segmented or segmented as other structure	Tissue class defectively segmented, substantially misclassified
3	50-75% segmented, 50-25% not segmented or segmented as other structure	Tissue is mostly segmented and scarcely misclassified
4	75-100% segmented, 25-0% not segmented or segmented as other structure	Tissue is accurately segmented with only minor misclassification
5	All lesions segmented, well defined and matching the shape of the lesions	Tissue is completely segmented, without misclassification

Abbreviations: WML = white matter lesions; WM = white matter; GM = Grey Matter; CSF = cerebral spinal fluid

2.1.2.3 Evaluation of Segmentation Approaches

All tissue types from the 7T segmentation algorithms were quantitatively assessed using 3T pseudo-labels as a reference. As described earlier, 3T pseudo-labels were created from up-sampled C-DEF 3T segmentation outputs that match 7T resolution ($0.7mm$ isotropic). Skull-stripping performance was evaluated using Bland-Altman analysis. The Dice similarity coefficient (DSC) for each segmentation class (whole brain, GM, WM, CSF, and WML) was calculated as described by Zou et al. (2004). The volumetric agreement for WML segmentations was also assessed.

The performance of PLAn 7T and nnU-Net 7T in WML segmentation was further validated against manually drawn lesion volumes on eight 7T scans. Volumetric agreement was calculated, and overlap measures (DSC, voxel, and

cluster level false positive ratio and false negative ratio, detection error rate, outline error rate) were extracted using FSL's *bianca_overlap_tools*, with definitions for all overlap metrics referenced in the original paper (Griffanti et al., 2016).

Reproducibility of nnU-Net 7T and PLAn 7T segmentations was evaluated in the 7T test–retest cohort. Bland–Altman analysis was used to estimate the percent error in brain, GM, WM, CSF, and WML segmentations. Volumetric agreement between sessions was further quantified using the intraclass correlation coefficient (ICC).

2.1.2.4 Statistical Analysis

All statistical analyses were performed using PRISM (version 9.3.1). Normality was assessed using the Shapiro–Wilk test. For group comparisons, repeated-measures ANOVA with Dunnett's correction for multiple comparisons was used to compare segmentation methods against the reference standard (3T pseudo-labels or manual annotations, as appropriate). Bland–Altman analysis was used to evaluate reproducibility and bias in volumetric estimates. Percent error was calculated as the absolute difference between segmentation volumes divided by their mean volume, with 95% limits of agreement (LoA) reported (Bland and Altman, 2003). Volumetric correspondence between segmentations was quantified using Spearman's correlation, as the data were not normally distributed. ICC was calculated using a two-way mixed-effects model with absolute agreement. A corrected p-value ≤ 0.05 was considered statistically significant. Unless otherwise stated, results are reported as mean $\pm$ standard deviation (SD).

2.1.3 Results

2.1.3.1 Qualitative Evaluation: Paired Cohort 1

Qualitative evaluation of all 7T segmentation algorithms was performed on the testing data from Paired Cohort 1 ($n = 20$) to compare their performance across brain tissues and WML. FreeSurfer 7T skull-stripping completely failed in 3 cases. These cases were removed from the whole brain, GM, WM, and CSF qualitative and quantitative analysis.

Table 2.2 shows the average qualitative ratings of all segmentation classes for FreeSurfer 7T, C-DEF 7T (which uses AFNI's *3dSkullStrip* for whole brain

segmentation), nnU-Net 7T, and PLAn 7T. nnU-Net 7T and PLAn 7T outperformed C-DEF 7T and FreeSurfer 7T in qualitative ratings across all segmentation classes (Table 2.2). nnU-Net 7T and PLAn 7T's improved whole brain segmentation, cortical boundary delineation, and WML segmentation can be seen in Figure 2.2 and Figure 2.3.

Table 2.2: Average blinded qualitative ratings across raters (mean, 95% CI) of Paired Cohort 1 testing dataset. Adapted from (Donnay et al., 2023)

Segmentation	**FreeSurfer 7T**	**C-DEF 7T**	**nnU-Net 7T**	**PLAn 7T**
Whole Brain	1.82 [2.07, 1.56]	4.03 [4.20, 3.86]	4.42 [4.51, 4.33]	4.40 [4.50, 4.30]
CSF	1.58 [1.86, 1.30]	2.76 [2.98, 2.55]	3.89 [4.08, 3.70]	3.90 [4.09, 3.71]
GM	1.95 [2.14, 1.75]	2.39 [2.63, 2.16]	3.34 [3.5, 3.18]	3.39 [3.55, 3.24]
WM	2.50 [2.79, 2.21]	3.16 [3.39, 2.93]	3.92 [4.01, 3.83]	3.92 [4.01, 3.83]
WML	1.95 [2.22, 1.67]	1.63 [1.90, 1.37]	3.74 [3.98, 3.49]	3.92 [4.16, 3.68]

Abbreviations: CSF = cerebral spinal fluid; GM = Grey Matter; WM = white matter; WML = white matter lesions

Expert raters rated nnU-Net 7T and PLAn 7T as the best methods for skull stripping, with mean scores of 4.42 (CI: 4.33, 4.51) and 4.4 (CI: 4.3, 4.5), respectively. In contrast, C-DEF 7T received a lower mean rating of 4.03 (CI: 3.7, 4.2), reflecting poorer segmentation in dorsal (Figure 2.2A) and infratentorial (Figure 2.2C) areas of the brain (Figure 2.2A) compared to nnU-Net 7T and PLAn 7T. A mean rating score of just 1.82 (CI: 1.6, 2.1) for FreeSurfer 7T reflected its low accuracy in skull stripping.

Expert ratings demonstrated nnU-Net 7T and PLAn 7T as the best methods for CSF segmentation, with scores of 3.9, compared to 1.58 for FreeSurfer 7T and 2.76 for C-DEF 7T (Table 2.2). nnU-Net 7T and PLAn 7T also outperformed other methods in distinguishing deep GM structures like the thalamus and globus pallidus (Figure 2.3B). Overall, both methods received the highest expert ratings for GM and WM segmentation, highlighting their superior performance across all tissue types. Between nnU-Net 7T and PLAn 7T, expert ratings (Table 2.2) indicated a minor improvement in GM segmentation (1.5%) and no change for CSF or WM in PLAn 7T relative to nnU-Net 7T.

nnU-Net 7T showed deficiencies in lesion detection, such as under segmentation of large lesions (Figure 2.3B) or missing small lesions (Figure 2.3D), despite outperforming other 7T methods. PLAn 7T was able to address these

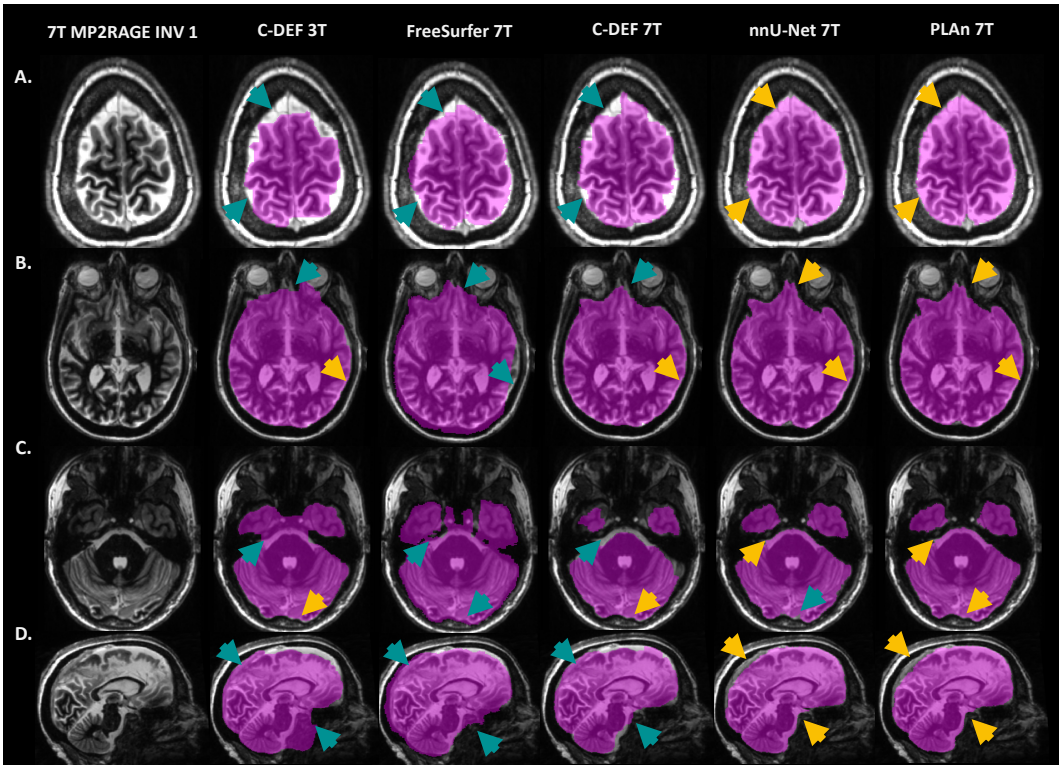


Figure 2.2: Examples of whole brain segmentations of 3T C-DEF reference and 7T segmentation algorithms. Representative axial (A–C) and sagittal (D) slices show skull-stripping outputs (magenta overlay) from various participants. Gold arrows highlight accurate skull-stripping performance, while teal arrows indicate errors. MP2RAGE inversion 1 is shown here for visualization purposes. As described in the text, MP2RAGE inputs for skull stripping across methods are as follows: inversion 1 for C-DEF models; T1w for FreeSurfer; and inversion 1, inversion 2, and T1map for both nnU-Net and PLAn. Image adapted from Donnay & Dieckhaus 2023 - originally created by H.D. and edited by C.D.

shortcomings and had a 5% increase in expert lesion ratings (3.92, CI: 3.68, 4.16) compared to nnU-Net 7T (3.74 CI: 3.49, 3.98).

2.1.3.2 Quantitative Evaluation: Paired Cohorts 1 and 2

Figure 2.4 shows the results of quantitative evaluations using Paired Cohorts 1 and 2. Using 3T pseudo-labels as reference segmentations, all 7T algorithms were quantitatively evaluated on Paired Cohort 1. Since these pseudo-labels were also used in pre-training PLAn 7T, this analysis was biased for PLAn 7T (biased analysis indicated with * in Figure 2.4), so an additional quantitative evaluation was performed using Paired Cohort 2 for nnU-Net 7T and PLAn 7T.

Bland-Altman plots showed that FreeSurfer 7T overestimated whole brain volume, producing a mean bias of 10% (CI: 3%, 19%) compared to the reference 3T pseudo-labels, while C-DEF 7T consistently underestimated whole brain volume, with a mean bias of 3.5% (CI: -8.2%, 1.1%) (Figure 2.4A). In contrast,

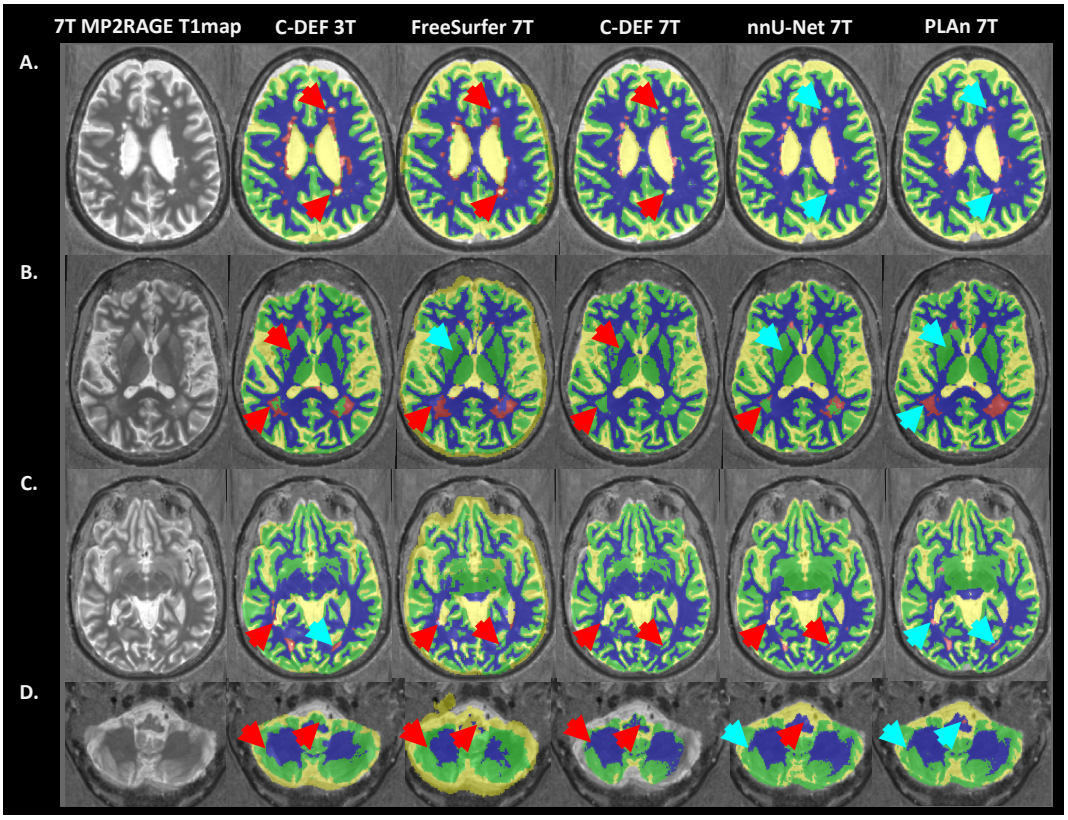


Figure 2.3: Representative axial slices illustrate the performance of different tissue segmentation methods (indicated at the top, including method and field strength) across several brain regions: (A-C) supratentorial and D) infratentorial, overlaid on 7T MP2RAGE T1map images. Blue arrows highlight well-segmented areas, while red arrows point to segmentation errors. In the segmentation overlays, blue represents WM, yellow for CSF, green for GM, and magenta for lesions. Image adapted from Donnay & Dieckhaus 2023 - originally created by H.D. and edited by C.D.

both nnU-Net 7T and PLAn 7T exhibited very low mean whole brain volume biases in Paired Cohort 1, with nnU-Net 7T at 0.88% (CI: -3.8%, 5.6%) and PLAn 7T at 0.89% (CI: -3.7%, 5.5%) (Figure 2.4A). However, in Paired Cohort 2, they showed higher bias (Figure 2.4B: nnU-Net 7T bias=5.8% CI:1.8%, 9.9%; PLAn 7T bias=6.2% CI:1.8%, 10.7%). C-DEF 7T achieved the highest overall mean whole brain DSC compared to C-DEF 3T, with a mean DSC of 0.96 (CI: [0.95, 0.97]). However, this was only significantly different from FreeSurfer 7T, which had statistically significantly lower DSC scores than all other methods (Figure 2.4C). C-DEF 7T showed similar skull-stripping results to C-DEF 3T in some brain regions, like the dorsal areas, but C-DEF 7T under-segmented infratentorial regions (Figure 2.2).

nnU-Net 7T and PLAn 7T had high DSC across all CSF, GM, WM, and WML (Figure 2.4D-G). FreeSurfer 7T, on the other hand, struggled with temporal lobe and cerebellum segmentation, mislabelling sulcal grey matter as CSF

(Figures 2.3C, D), resulting in the lowest DSC (mean = 0.65, CI: 0.63, 0.69, Figure 4D). C-DEF 7T showed better CSF segmentation but still had a lower DSC (0.7, CI: 0.67, 0.73) compared to nnU-Net 7T's DSC of 0.72 (CI: 0.69, 0.75).

In Paired Cohort 1, PLAn 7T had small but statistically significant improvements in CSF (1%), GM (0.6%), and WM (0.5%) segmentation compared to nnU-Net 7T (Figure 2.4). In Paired Cohort 2, PLAn 7T only had higher DSC in GM and WM. The quantitative results confirmed improved WML in PLAn 7T compared to other 7T algorithms, including nnU-Net 7T. PLAn 7T demonstrated significantly higher DSC than all 7T algorithms and slightly better volume correlations than nnU-Net 7T (Figure 2.4G, H).

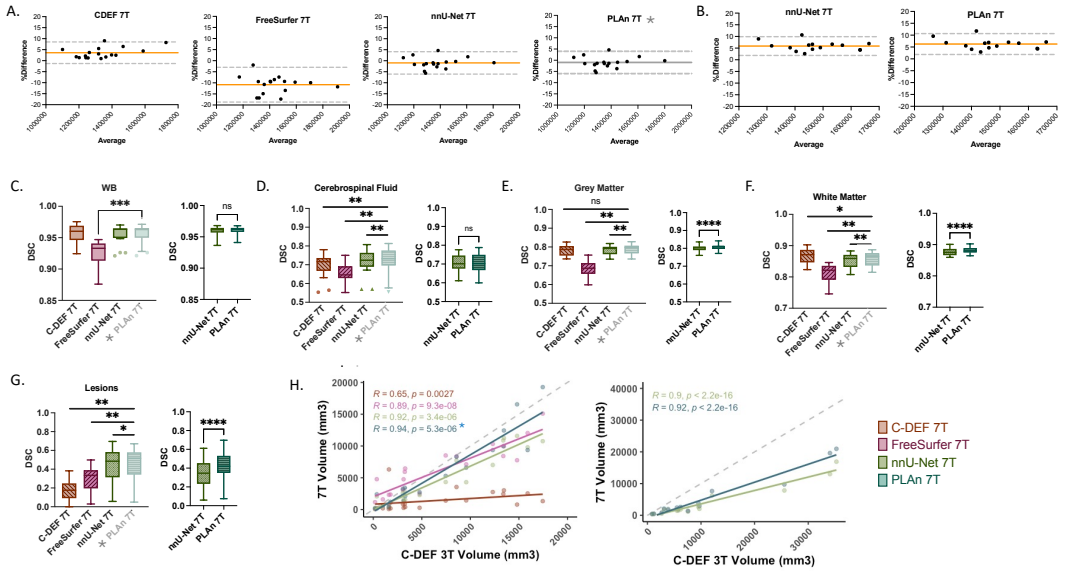


Figure 2.4: Quantitative evaluation of 7T segmentation algorithms compared to 3T Reference. A) Bland-Altman plots of total whole brain volume calculated from 7T methods in compared to C-DEF 3T pseudo-labels using Paired Cohort 1. B) Bland-Altman plots of whole brain volume calculated from nnU-Net 7T and PLAn 7T compared to C-DEF 3T pseudo-labels using Paired Cohort 2. Dice Similarity Coefficient (DSC) comparing all 7T algorithms to the C-DEF 3T pseudo-labels using Paired Cohort 1 (left graph, $n = 20$) and comparing nnU-Net 7T and PLAn 7T using Paired Cohort 2 (right graph, $n = 15$) for C) WB, D) CSF, E) GM, F) WM, and G) WML. H) Correlation plots of WML volume from 7T segmentation methods vs C-DEF 3T pseudo-labels using Paired Cohort 1 (left graph, $n = 20$) and from nnU-Net 7T and PLAn 7T (right, $n = 15$) vs the C-DEF 3T pseudo-label reference, with identity (dashed grey) and regression (solid coloured) lines. Significant differences compared to PLAn 7T are indicated with asterisks. Grey * indicate comparisons that are biased.

2.1.3.3 Quantitative Evaluation: 7T Validation Cohort

In an unseen 7T cohort, the lesion segmentation performance of PLAn 7T and nnU-Net were assessed using manually annotated lesion segmentations as

a gold standard. PLAn-7T showed enhanced lesion segmentation across various lesion types, including diffuse WML and periventricular lesions, which nnU-Net 7T frequently misclassified as GM, WM, or CSF (Figures 2.5A-D). Quantitative analysis revealed a significantly higher DSC for PLAn 7T (mean = 0.86, CI: [0.83, 0.90]) compared to nnU-Net 7T (mean = 0.72, CI: [0.78, 0.67], $p < 0.0001$). Additional overlap metrics using FSL's *bianca_overlap_measures* confirmed the superiority of PLAn 7T over nnU-Net (Appendix B.1). Furthermore, PLAn 7T exhibited a stronger slope (slope = 0.72), closer to unity than nnU-Net 7T (slope = 0.53), which systematically under segmented lesions at higher lesion volumes (Figure 2.5F).

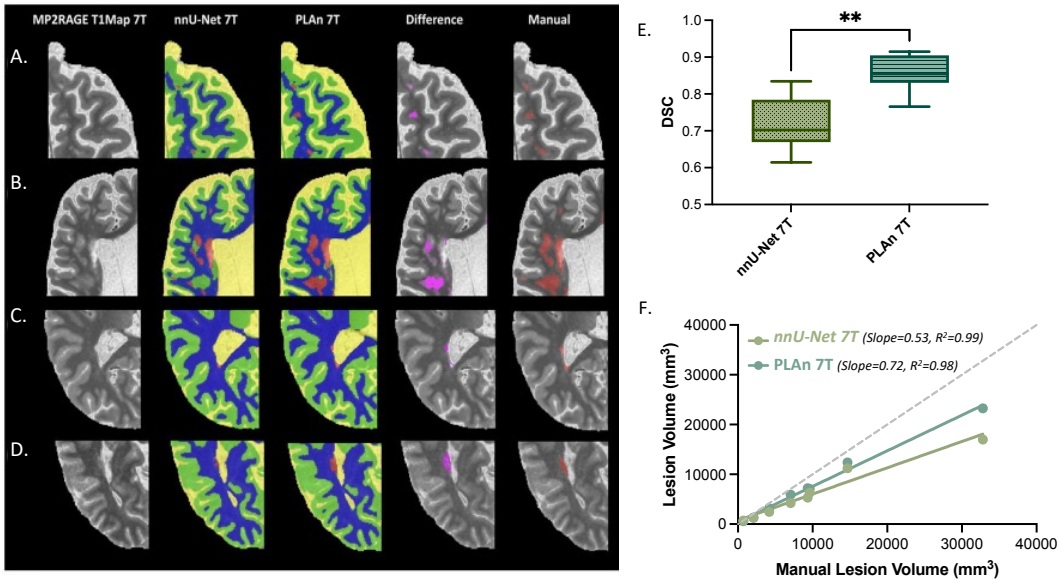


Figure 2.5: Comparison of lesion segmentation performance between PLAn 7T and nnU-Net 7T on an unseen cohort. This figure presents segmentation results from nnU-Net (column 2) and PLAn (column 3) across different participants, focusing on A) punctate lesions in the frontal lobe, B) periventricular and deep white matter lesions, and C), D) lesions in the occipital horn of the lateral ventricle. Segmented lesions (in red) are overlaid on the 7T MP2RAGE T1 map. A voxelwise subtraction of lesion masks from nnU-Net 7T and PLAn 7T is presented in pink in column 4. E) The Dice similarity coefficient (DSC) for nnU-Net and PLAn using manual 7T segmentation masks, are compared ($p < 0.0001$). F) Lesion volume measurements produced by nnU-Net (green) and PLAn (blue) are highly correlated with manual segmentation mask volume.

2.1.3.4 Reproducibility Evaluation: Test-Retest Cohort

The results of the reproducibility test for all tissue types segmented with PLAn 7T and nnU-Net 7T from 8 participants are presented in Figure 2.6. Both methods demonstrated high reproducibility of total tissue volumes between test-retest sessions. Specifically, the mean bias for GM, WM, CSF, and WB volumes was less than 1% for both PLAn 7T and nnU-Net 7T, with confidence intervals consistently under 5%. PLAn 7T exhibited lower variability

in WM and slightly lower variability in WB and lesions compared to nnU-Net 7T. Conversely, nnU-Net showed lower variability in CSF and slightly lower variability in GM. Additionally, all tissue volumes were significantly correlated (Table 2.3), and the average percentage error across all tissue types was less than 2%.

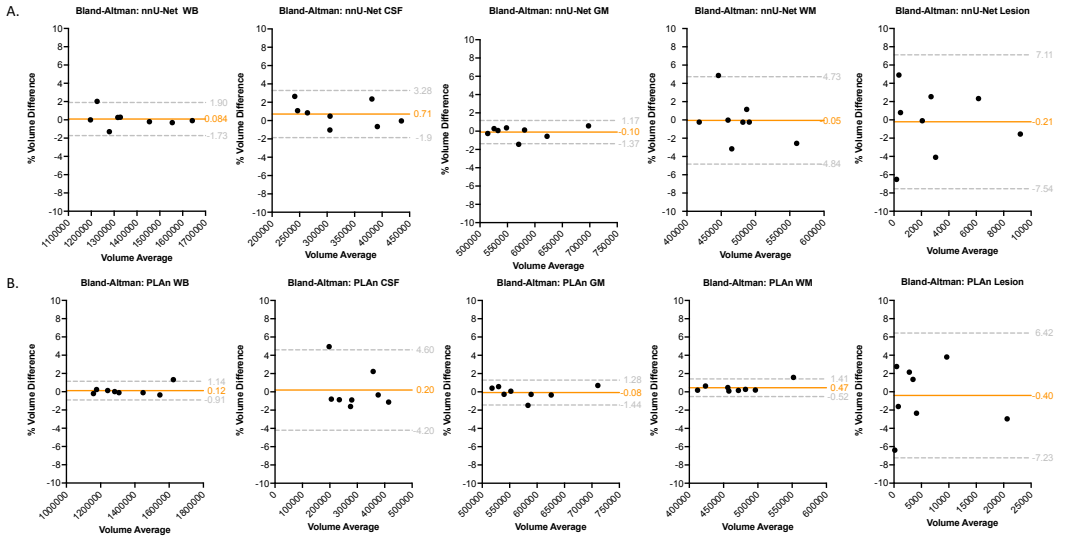


Figure 2.6: Evaluation of the reproducibility of PLAn 7T and nnU-Net 7T in a test-retest dataset. Bland-Altman plots comparing the difference in tissue class volumes segmented with A) nnU-Net 7T and B) PLAn. Participants ($n = 8$) were scanned twice using the same scanner and protocol. The percentage difference between sessions is calculated as: % Difference = 100 × [(Session 2 - Session 1) / Average]. The mean bias (dashed orange line) and 95% limits of agreement (dashed grey lines) are indicated. Each point represents the volume for a tissue class from one participant.

2.1.4 Discussion

In this section, I describe the development and evaluation PLAn 7T, a comprehensive whole-brain and lesion segmentation tool for 7T MRI data. This model integrates whole-brain, brain tissues, and WML segmentation into a single, unified process. The approach leveraged nnU-Net, a publicly available medical image segmentation method that has strong segmentation performance even with a limited number of well-annotated training examples. By incorporating pre-training on 3T pseudo-labels into nnU-Net 7T (as part of PLAn 7T), lesion segmentation performance was significantly enhanced. This pre-training step leverages the abundance of 3T pseudo-labels to overcome the challenge of limited 7T training data, ultimately improving segmentation accuracy for lesions in MS patients.

PLAn 7T and nnU-Net 7T outperformed C-DEF 7T and FreeSurfer 7T across all tasks at 7T. nnU-Net's success in skull stripping aligns with prior

Table 2.3: Volumetric evaluation of all tissue classes between sessions in the test-retest dataset using nnU-Net 7T and PLAn 7T

Algorithm	**Tissue**	**Spearman Correlation**		**Average Percent Error**		**ICC**		
		ρ	**p-value**	**Avg. % Error**	**SD**	**ICC**	**ICC Lower**	**ICC Upper**
nnU-Net 7T	WB	1	<0.0001	0.56	0.71	1.00	1.00	0.99
PLAn 7T	WB	1	<0.0001	0.31	0.42	1.00	1.00	1.00
nnU-Net 7T	GM	1	<0.0001	0.46	0.43	1.00	0.99	1.00
PLAn 7T	GM	1	<0.0001	0.51	0.43	1.00	0.99	1.00
nnU-Net 7T	CSF	0.98	<0.0001	1.14	0.90	1.00	0.99	1.00
PLAn 7T	CSF	1	<0.0001	1.60	1.47	1.00	0.99	1.00
nnU-Net 7T	WM	0.98	<0.0001	1.56	1.78	0.97	0.84	0.99
PLAn 7T	WM	1	<0.0001	0.45	0.49	1.00	0.98	1.00
nnU-Net 7T	WML	1	<0.0001	2.85	2.17	1.00	1.00	1.00
PLAn 7T	WML	1	<0.0001	2.92	1.60	1.00	1.00	1.00

Abbreviations: WB = Whole Brain; CSF = cerebral spinal fluid; GM = Grey Matter; WM = white matter; WML = white matter lesions; ICC = intra-class correlation; SD = standard deviation

studies that established the efficacy of U-Net-derived convolutional neural networks (CNNs) for brain extraction, including in MS patients (de Oliveira et al., 2022; Hwang et al., 2019; Wang et al., 2021). In contrast, FreeSurfer and C-DEF struggled with under- or over-skull stripping, which impacted FreeSurfer's downstream segmentation tasks. The poor skull stripping in FreeSurfer significantly reduced the quality of CSF segmentation, highlighting the importance of accurate skull stripping in maintaining segmentation quality. By integrating skull stripping and segmentation into a single step, PLAn and nnU-Net 7T demonstrated a more efficient and effective approach, with nnU-Net 7T combining the strengths of C-DEF 7T (good CSF segmentation and cortical boundary delineation) and FreeSurfer (detailed GM segmentation). However, FreeSurfer's performance in cortical segmentation, especially in the temporal lobe and cerebellum, suffered from segmentation errors, possibly due to strong 7T bias fields and cases with signal dropout in the infratentorial brain. This reduction in segmentation accuracy was systematic and localized, consistent with previous studies' findings that noted difficulties in correcting bias fields at ultra-high field strengths (Haast et al., 2018).

nnU-Net demonstrated good lesion segmentation at 7T despite being trained

with only five labels, a much smaller dataset than the typical datasets used in MRI segmentation models. For PLAn 7T, pre-training was performed using the up-sampled 3T segmentation outputs as pseudo-labels. Although these pseudo-labels were imperfect, including errors in skull stripping from C-DEF at 3T, PLAn 7T still yielded high-quality segmentation results comparable to the default nnU-Net 7T. This suggests that PLAn's fine-tuning successfully corrected errors in the 3T pseudo-labels while preserving the optimal features of the 7T baseline model, such as detailed GM structures and cortical boundaries. Recent studies have employed automated or semi-automated pseudo-labels followed by few-shot fine-tuning, suggesting that even small datasets (1-3 images) can be sufficient for fine-tuning, though smaller structures, such as the amygdala and accumbens, benefit from additional data (Bermudez et al., 2020; Kushibar et al., 2019; Svanera et al., 2021). Further research could investigate whether using a larger pre-training dataset would significantly improve segmentation.

FLAIR imaging often plays an important role in 3T segmentation algorithms for detecting WML. FLAIR uses an inversion recovery pulse sequence to suppress signals from fluids. This sequence uses a 180-degree radio-frequency pulse that inverts the magnetization of tissue protons, followed by a long inversion time (TI), where the longitudinal magnetization begins to recover. At the selected TI, the signal from fluid remains close to zero and is suppressed. As a result, CSF, which has high water content and is characterized by long T1 relaxation times, is nulled and remains dark in the final image. WMLs, which have shorter T1 and T2 times, appear as hyperintense regions, which have high contrast against the dark CSF and the intermediate signal of WM. MP2RAGE does not include CSF suppression, and because the T1-weighted signal is less sensitive to free water, WML and CSF intensities overlap more. Thus, the T2-FLAIR contrast offers better intensity differentiation between lesions with CSF and GM than the MP2RAGE contrast.

The absence of FLAIR contrast at 7T may contribute to why C-DEF 7T fails to detect lesions while C-DEF 3T does not. At 3T, C-DEF has been shown to outperform U-Net-based segmentation algorithms when trained on limited labels (Dieckhaus et al., 2022). However, nnU-Net 7T clearly outperformed C-DEF 7T, suggesting that MP2RAGE images alone do not provide sufficient contrast for C-DEF to reliably separate lesions from surrounding tissues at 7T. Since C-DEF extracts local image statistics that are highly contrast-dependent, the inclusion of FLAIR contrasts may significantly enhance its performance.

Conversely, nnU-Net is less dependent on specific contrasts as it relies more on morphological features and contextual information from both local and global regions. Furthermore, C-DEF's feature extraction process is influenced by the number of contrasts available during training; thus, incorporating multiple contrasts (such as FLAIR) provides more training features. PLAn 7T showed superior performance to nnU-Net 7T in identifying diffuse and periventricular lesions. In particular, PLAn 7T demonstrated the ability to detect subtle lesion features associated with 3T FLAIR hyperintensities, such as large, diffuse WML, even in the absence of 7T FLAIR. This ability likely stems from the pre-training dataset labels created using C-DEF 3T, which incorporates FLAIR as one of its modalities.

There are limitations to this study. Segmentation algorithms were primarily assessed in an off-the-shelf fashion, with little or no hyper-parameter tuning or modification. While dedicated fine-tuning of methods may improve results for specific cohorts, one of the reasons the nnU-Net framework was chosen was to avoid the need for extensive fine-tuning, making it suitable for potential wider clinical application.

Additionally, C-DEF 3T segmentations were used as a reference scan when comparing all segmentation methods. However, these cannot be treated as a perfect 7T gold standard due to inherent differences between 3T and 7T segmentations. Previous research has shown that lesion segmentation volumes derived from FLAIR contrasts tend to be higher than those derived from MP2RAGE scans, though the resulting volumes are highly correlated (Spini et al., 2020). In line with this, C-DEF 3T reference lesions exhibited higher lesion volumes than 7T methods that did not include FLAIR in their processing pipeline. To address this limitation, blinded neurologist ratings that were not based on 3T segmentations were included, and performance on an unseen 7T cohort with manual lesion segmentations and test-retest data were also evaluated.

In conclusion, this study has laid the groundwork for developing a robust whole-brain and lesion segmentation tool for 7T MRI data using the MP2RAGE modality. While efforts are underway to extend this research to assess the generalizability of PLAn across different datasets and contrasts, such as T2*-GRE and FLAIR, for the purposes of this thesis, the analysis focused on developing a segmentation tool using a single modality, which offers substantial benefit with respect to scan time required. This tool will significantly contribute to the

thesis's broader aims, including the analysis of longitudinal changes in chronic WML.

2.2 Impact of Longitudinal Analysis Pipelines Quantitative T1 and Volume

2.2.1 Introduction

T1 mapping is a valuable biomarker for distinguishing lesions from NAWM in MS. The sensitivity of qT1 to pathological changes offers critical insights into disease severity, disability progression, and cognitive decline, making it a valuable tool in clinical practice and research. Despite the lack of consensus on the optimal T1 mapping sequence (Ropele et al., 2011; Taylor et al., 2016), the MP2RAGE sequence produces high-quality qT1 maps that are sensitive in detecting MS lesions with clinically compatible scan times (Granziera et al., 2021; Marques et al., 2010). However, understanding the test-retest variability of qT1 measurements in MS cohorts is essential, particularly in the context of longitudinal studies and clinical trials, where reliable and consistent measurements are crucial for assessing disease progression and treatment effects.

The reproducibility of qT1 measurements is influenced by various factors, including hardware, software, and sequences (Bane et al., 2018). Nevertheless, studies have shown that the MP2RAGE sequence offers high reproducibility of qT1 measurements across different sites and healthy participants at both 3T and 7T (Rahmanzadeh, Weigel, Lu, Melie-Garcia, Nguyen, Cagol, La Rosa, Barakovic, Lutti, Wang, Bach Cuadra, Radue, Gaetano, Kappos, Kuhle, Magon and Granziera, 2022; Voelker et al., 2016, 2021) and even in spinal cord imaging (Rasoanandrianina et al., 2019). For example, a single-vendor, multisite study showed strong reproducibility of 7T MP2RAGE qT1 maps, particularly in the infratentorial region (Voelker et al., 2016, 2021). In this study, inter-site volume and qT1 coefficients of variation (CoV) were calculated for various tissues in two healthy participants. In GM and WM, the CoV was below 2% for volume and under 1% for qT1 values. However, the cerebellum exhibited significantly higher qT1 CoV values (9.3% and 15.4% for each participant) due to lower signal levels in inferior brain regions, which led to more frequent segmentation and measurement failures (Voelker et al., 2016). Additionally, in a follow-up study, more homogenous site configurations yielded

higher inter-site reproducibility, compared to sites with more heterogeneous setups of key imaging components, such as magnet design and gradient coils (Voelker et al., 2021). Despite these promising findings in healthy participants, there remains a notable gap in the literature regarding the reproducibility of 7T MP2RAGE qT1 maps in MS patients, particularly in relation to MS lesions.

Image processing pipelines also impact the reliability of MRI biomarkers. Several studies have explored the reliability of image processing tools and approaches for quantifying volumetric measurements of the whole brain, cortical, and subcortical structures, particularly in diseases like Alzheimer's (Jovicich et al., 2013; Mulder et al., 2014; Okubo et al., 2016; Teipel et al., 2008). However, to our knowledge, no research has specifically examined the effects of image processing pipelines on qT1 measures in the context of MS lesions and normal appearing tissues in MS patients. Reliable biomarkers are essential for correctly interpreting trial endpoints and longitudinal studies, since variability introduced by the pipeline could obscure true biological change. It is also important to assess whether reproducibility depends on tissue characteristics, as such systematic influences could bias interpretation. This study aims to investigate the stability of qT1 measures in MS patients using the 7T MP2RAGE sequence after MRI scanner re-localization. I evaluate the impact of various longitudinal image analysis pipelines on test-retest variability of qT1 in different tissues, as well as on test-retest of lesion volume measurements. The aim is to identify which measures and pipelines show the highest reproducibility in order to inform the choice of approaches to take forward into clinical trials.

2.2.2 Methods

This section uses a 7T Test–Retest Cohort (Appendix A.2) acquired as part of a natural history protocol (Appendix A.1).

2.2.2.1 Overview of Analysis Pipelines

Inversion 1, Inversion 2, and T1w images from the MP2RAGE scan were bias-field corrected. All test-retest cases were segmented using PLAn 7T (Chapter 2.1) for use in downstream image processing. Five distinct image processing pipelines were applied to all participants (Figure 2.7) to obtain whole brain, GM, WM, CSF, and WML segmentations. The pipelines were chosen to represent a spectrum of approaches ranging from fully automated methods to manual "gold standard" approaches. Pipeline 1 provides a simple, fully automated

strategy in which consistent labels are maintained across sessions through registration. Pipeline 2 adopts a halfway-space registration approach, which reduces bias introduced by lesion segmentation differences. Pipeline 3 applies PLAn segmentations independently to each session, representing a standard automated analysis workflow. Pipeline 4 incorporates manual editing of lesions in addition to PLAn segmentation, reflecting the gold standard practice of using expert-corrected lesion masks. Finally, Pipeline 5 combines automated lesion registration with a boundary-correction method, addressing limitations of other automated pipelines by enabling consistent lesion tracking across sessions while refining lesion boundaries.

Pipeline 1 - Registration of images and masks from session 1 to session 2: PLAn 7T segmentation was applied to both sessions separately to acquire brain masks. The brain masks were multiplied by T1w images to create brain-extracted T1w images. The session 1 brain-extracted T1w image was rigidly registered (6 degrees of freedom, DoF) to the session 2 brain-extracted T1w image using FMRIB's linear registration tool (Jenkinson and Smith, 2001). The PLAn segmentation output for session 1 was split into four masks: GM, WM, CSF, and total WML. The total WML mask was spatially split into individual lesion clusters using FSL's *cluster* tool (degree of connectivity = 6). These masks were used to extract qT1 metrics and volume from the session 1 qT1 map. Then, the registration matrix was applied to session 1 GM, WM, CSF, and individual lesion masks to move masks to session 2 space, using nearest-neighbour interpolation. The registered masks were used to extract qT1 metrics and volume from the session 2 qT1 map.

Pipeline 2 - Halfway registration of images and masks: The PLAn 7T segmentation output from both sessions was split into four masks: GM, WM, CSF, and total WML. For both sessions, the total WML mask was spatially split into individual lesion clusters using FSL's *cluster* tool (degree of connectivity = 6). The segmentation labels for individual WML in session 2 were matched to session 1 (described below). Session 1 and session 2 brain-extracted T1w images were rigidly registered into halfway space using a registration code adapted from SIENA (Smith et al., 2001). The corresponding registration matrices were applied to both sessions' qT1 maps using sinc interpolation and to the segmentation masks using nearest-neighbour interpolation. The intersection of the session 1 and session 2 segmentation masks was used to extract qT1

measures from qT1 maps. Only lesions labelled in both session masks were included.

Pipeline 3 - Independent PLAn 7T segmentation: The PLAn 7T segmentation output from both sessions was split into four masks: GM, WM, CSF, and total WML. For both sessions, the total WML mask was spatially split into individual lesion clusters using FSL's *cluster* tool (degree of connectivity = 6). Segmentation labels for individual WML in session 2 were matched to those in session 1. For each session, qT1 measures and volume were extracted from each segmentation.

Pipeline 4 - Independent PLAn 7T segmentation plus manual editing: GM, WM, and CSF were processed the same way as described in pipeline 3. However, unlike pipeline 3, the total WML segmentation mask was manually edited in both sessions. Manually editing WML masks generated from automated segmentation outputs is commonly referred to as gold standard segmentation in MS (Wattjes et al., 2021). Segmentation labels for individual WML in session 2 were matched to those in session 1. For each session, qT1 measures and volume were extracted from each segmentation.

Pipeline 5 - Registration of individual WML masks plus boundary correction: The PLAn 7T segmentation output from both sessions was split into four masks: GM, WM, CSF, and total WML. For session 1, the total WML mask was spatially split into individual lesion clusters using FSL's *cluster* tool (degree of connectivity = 6). Brain masks from PLAn 7T were applied to the T1w images, which were linearly registered from session 1 to session 2 using FMRIB's linear registration tool (12 DoF). The registration matrix was applied to the individual lesion masks from session 1. Individual lesion segmentation masks for session 1 and registered individual lesion segmentation masks for session 2 both underwent a boundary correction pipeline (Figure 2.7C). Each lesion mask was dilated twice into the WM using a 0.7 mm^3 sphere kernel. Then, a bimodal Gaussian curve was fitted to the histogram of qT1 values from the dilated lesion mask, and the lesion mask was thresholded at the intersection of the two curves and binarised. The segmentation was refined by first removing stray voxels, where only the largest spatially connected cluster was retained, effectively removing smaller, disconnected voxels. Then, to fill any potential holes in the segmentation, the segmentation was split into individual z-axis slices, and FSL's *fslmaths* tool was used with the *-fillh* option to fill internal holes within each slice. After all lesions had been boundary corrected, they

were combined to create a final total lesion volume mask. Original PLAn GM, WM, and CSF masks were corrected using the final total lesion mask.

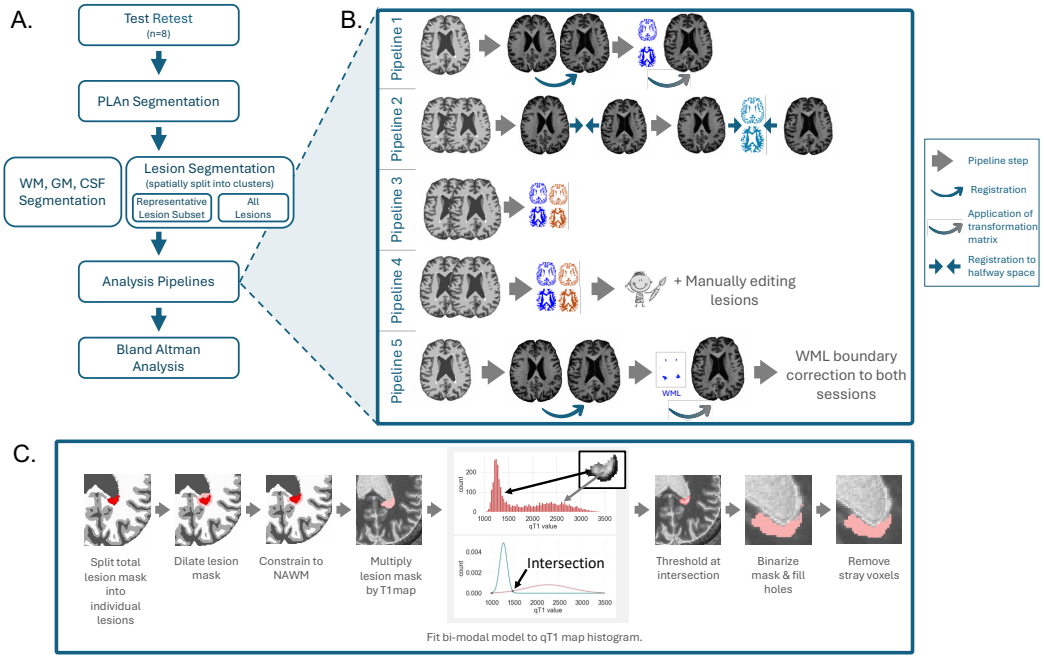


Figure 2.7: Schematic overview of 5 analysis pipelines used for image analysis. A) 7T MP2RAGE test-retest data was collected in eight participants. PLAn segmentation was run on all sessions. Test-retest variability was determined in GM, WM, CSF and lesions (first in a subset of lesions and then in all lesions) in different analysis pipelines using Bland-Altman Analysis. B) Overview of different analysis pipelines. C) Boundary correction steps that are applied to individual WML masks in Pipeline 5.

2.2.2.2 Test-Retest Variability of Quantitative T1 in Brain Tissues

For analysis pipelines 1, 2, 3, and 5, qT1 summary statistics (25^{th} percentile, median, 75^{th} percentile, and mean) were extracted for GM, WM, and CSF using FSL's *fslstats* tool. Quartile values were included because tissue damage and repair processes, such as demyelination and remyelination, are often partial and spatially heterogeneous within lesions. Damage is typically most severe at the lesion core, with partial repair more likely toward the periphery. As such, quartile-based measures may be more sensitive to subtle changes at the edges of lesions, where dynamic repair processes occur (Reich et al., 2015). While lesion sub-segmentation into core and edge regions could, in theory, provide similar insights, this approach is highly sensitive to segmentation inaccuracies due to partial voluming effects. In contrast, quartile-based measures offer a more robust and sensitive approach for capturing subtle peripheral changes without requiring precise delineation of lesion subregions.

Pipeline 4 was excluded from this analysis, as its processing for GM, WM, and CSF was identical to that of pipeline 3. The test-retest reliability of qT1 measurements across different analysis pipelines and summary statistics was assessed using Bland-Altman plots. For all Bland-Altman plots in this section, bias was calculated as the difference between measurements from session 1 and session 2, and the LoA were calculated as 1.95 times the standard deviation of the difference between session 1 and session 2.

2.2.2.3 Test-Retest Variability of Quantitative T1 and Volume in Lesions

The test-retest reliability of qT1 times and lesion volume for individual WML were evaluated in two stages.

Lesion Analysis Stage 1: First, 2-4 lesions per participant were manually identified to control for lesion selection across different pipelines. These lesions were chosen to represent a range of lesion types, considering both location and size while ensuring they represented the participant's overall lesion profile. This approach allowed for a fair comparison of the test-retest reliability of the same lesions across all pipelines.

Lesion Analysis Stage 2: Next, each analysis pipeline was applied to all lesions across all participants. Importantly, the number of lesions identified differs between pipelines. Pipelines 1 and 5 automatically match lesions between session 1 and session 2 during their registration process. In contrast, Pipelines 2, 3, and 4 require a lesion matching algorithm. For these pipelines, a rigid registration matrix was applied to the session 1 WML mask (with distinct labels for separate clusters) to align it with the session 2 WML masks. Then, the labels in session 2 were renamed based on the labels from session 1 that had the most overlap, ensuring that corresponding lesions across sessions were accurately identified. The lesion labels in session 2 were then updated based on the session 1 label with the greatest overlap, ensuring accurate identification of corresponding lesions across sessions.

For both lesion analyses, qT1 summary statistics (25^{th} percentile, median, 75^{th} percentile, mean, and standard deviation) were extracted along with lesion volume using FSL's *fslstats* tool. The test-retest reliability of qT1 measurements and volume across different analysis pipelines and for different summary statistics was assessed using Bland-Altman plots.

Exploratory analyses: To aid the interpretation of the results in the main analyses, I also examined whether lesion characteristics could account for differences in qT1 median or volume in both lesion analysis stages across different analysis pipelines. The percent of edge voxels for each lesion was calculated by dividing the number of voxels along the lesion boundary by the total number of lesion voxels (Figure 2.8). This metric provided insight into how segmentation boundary variability might affect reliability. Additionally, the qT1 standard deviation of individual lesions as a measure of intensity variability was included, to assess whether a lesion's distribution of intensities contributed to test-retest variability. For *Lesion Analysis Stage 1*, lesions were also qualitatively classified based on their location (brainstem, deep white matter (DWM), leukocortical, and periventricular) and based on their definition (well-defined or diffuse). The lesion definition was based on the sharpness of the lesion boundaries (Figure 2.8).

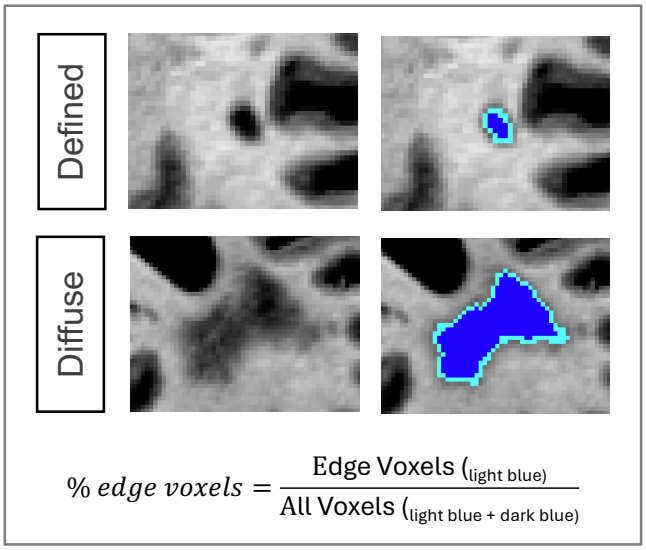


Figure 2.8: Example lesion characterized as *defined* (with a sharp lesion border) and *diffuse* with an indistinct lesion border. For both example lesions, edge voxels are highlighted in light blue.

2.2.2.4 Statistical Analysis

All Bland-Altman plots are shown in terms of bias and 95% LoA agreement. The LoA range refers to the absolute difference between the upper and lower LoA. The bias and LoA values for all Bland-Altman plots can be found in Appendix B.2.

2.2.3 Results

2.2.3.1 Test-Retest Variability of Quantitative T1 in Brain Tissues

The test-retest variability of qT1 times in GM, WM, and CSF is summarized in Figure 2.9. Figure 2.9A compares Bland-Altman plots of median qT1 across different analysis pipelines. Pipeline 2 showed the most stable results in GM, with no evident bias ($1 \pm 9.1ms$) and a range of LoA spanning 35.7ms. Pipelines 1, 3, and 5 exhibited higher ranges of LoA in GM ($40.5ms$, $52.5ms$, and $52.5ms$, respectively). Pipelines 2, 3, and 5 had comparable test-retest variability in WM and CSF. In WM, the bias and ranges of LoA were relatively low across pipeline 2 (bias=$0.6ms$, $11ms$), pipeline 3 (bias=$-1.3ms$, $13ms$), and pipeline 5 (bias=$-0.9ms$, $-13ms$). However, qT1 variability and bias were markedly higher in CSF compared to WM across all pipelines (Figure 2.9A, bottom).

Using pipeline 5, the test-retest reliability of various qT1 summary statistics was further explored, including the 25^{th} percentile, 75^{th} percentile, mean, and median for GM, WM, and CSF (Figure 2.9B). The bias was lowest for the median qT1 in GM and WM. In GM, the lowest LoA range was observed for the median qT1 ($55ms$). In contrast, the ranges of agreement for WM were comparable across all summary statistics, with the qT1 25^{th} percentile having the lowest LoA, followed closely by the qT1 median (Figure 2.9B). The 75^{th} percentile showed the least bias ($12.91ms$) in CSF compared to other summary statistics, which had biases between $15ms$ and $20ms$. The LoA were lowest for 75^{th} percentile in CSF, followed closely by the mean.

2.2.3.2 Test-Retest Variability of Quantitative T1 in White Matter Lesions

The test-retest variability of qT1 times and volumes in individual WMLs was first evaluated in 2-4 representative lesions per participant, resulting in 26 lesions. Figures 2.10A and 2.10B show Bland-Altman plots across different analysis pipelines for median qT1 and volume, respectively. Pipeline 3 exhibited the highest bias and LoA for both median qT1 (Figure 2.10A) and lesion volume (Figure 2.10B). Pipeline 2 demonstrated the lowest test-retest variability for median qT1. Test-retest variability for median qT1 was comparable between Pipeline 4 and 5, with Pipeline 5 showing a slightly narrower LoA by 17.11 ms (Figure 2.10A). However, Pipeline 4 had more stable volume measurements than Pipeline 5, with lower bias and narrower LoA (Figure 2.10B). The

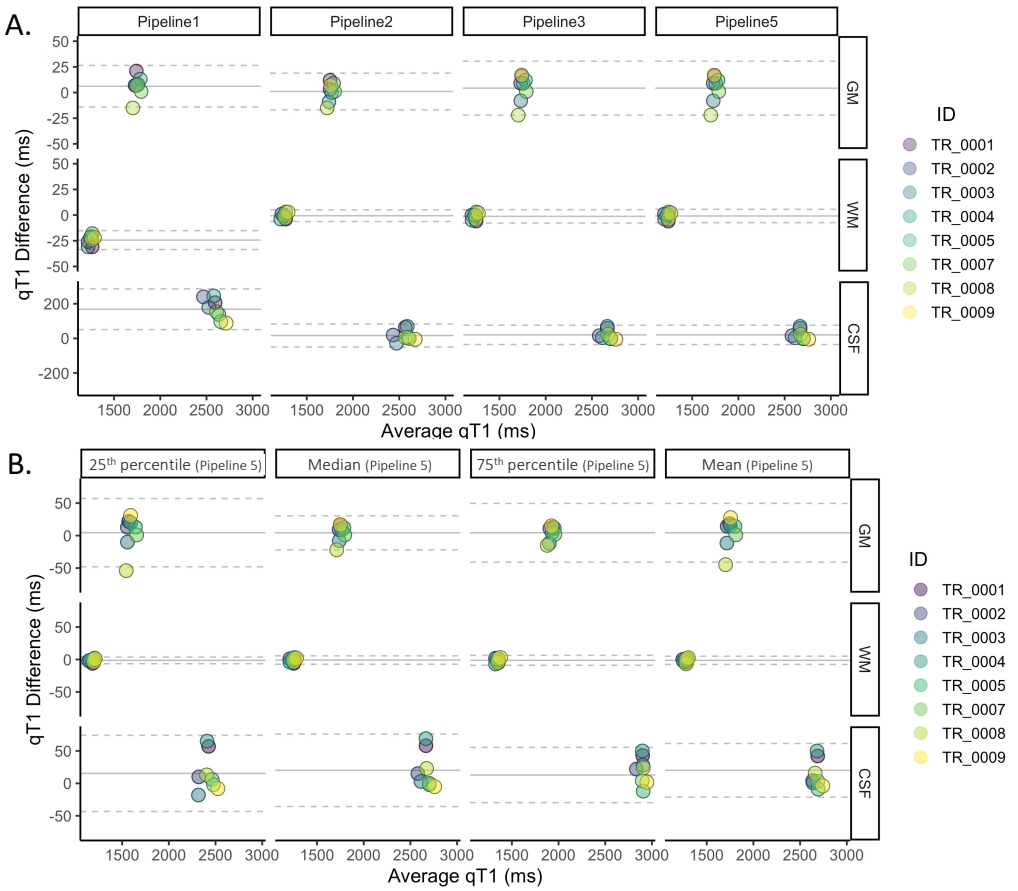


Figure 2.9: Within-scanner Test-Retest Variability of MP2RAGE qT1 Times in GM, WM, and CSF. A) Bland-Altman plots of median qT1 times across GM, WM, and CSF using different analysis pipelines. CSF Y-axis scale is different than GM and WM panels. B) Bland-Altman plots of qT1 summary statistics using Pipeline 5 across GM, WM and CSF. All Bland-Altman plots show the difference in qT1 measurements between session 1 and session 2 (y-axis) versus the average qT1 value (x-axis). The solid line indicates the bias (mean difference), while the dashed lines represent the 95% limits of agreement (LoA). Each point represents a participant in the test-retest cohort (n=8), with whole-brain 7T MP2RAGE scans. See Appendix B.2 for all Bland-Altman values.

comparison of qT1 summary statistics in pipeline 5 revealed that the median and 25$^{\text{th}}$ percentile qT1 were the most stable measurements, with the 25$^{\text{th}}$ percentile showing only a 5.8 ms narrower range of LoA than the median (Figure 2.10C).

The selected lesions were predominantly located in deep WM ($n = 16$), followed by periventricular ($n = 7$). Deep WM lesions showed a wide spread of qT1 variability, although most were clustered around low differences between session 1 and session 2 (Figure 2.11D). Three out of the five periventricular lesions showed low qT1 variability, while one had the highest variability across all selected lesions (Figure 2.11D, top). Leukocortical lesions ($n = 2$) exhibited slightly more qT1 variability than deep WM and periventricular lesions, but the

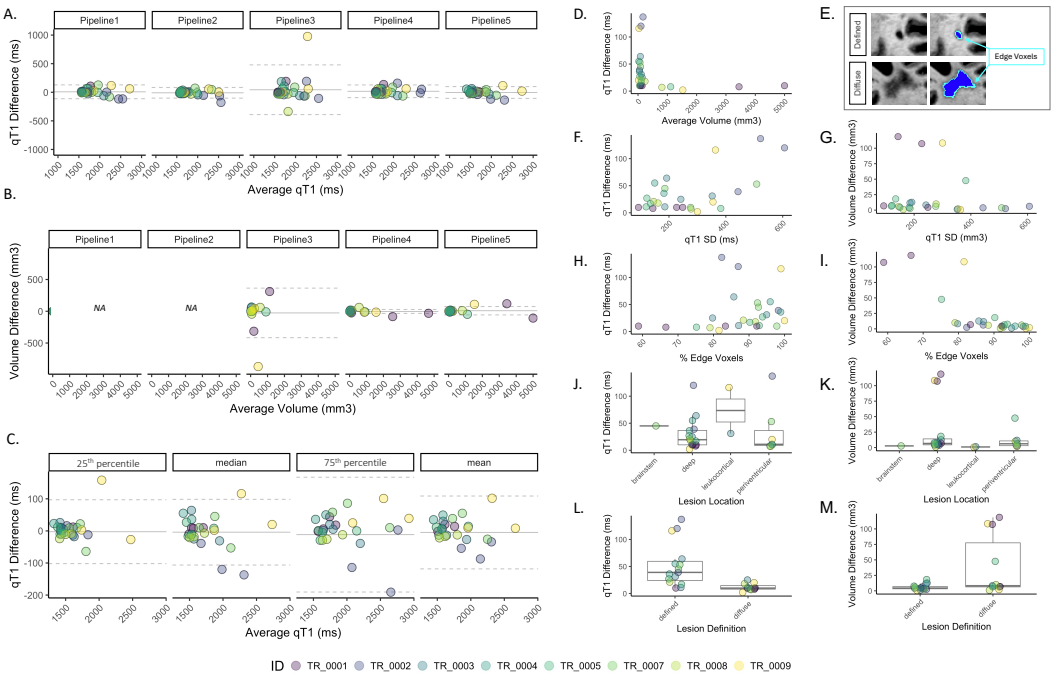


Figure 2.10: Within-scanner Test-Retest Variability of MP2RAGE qT1 times and volume in 27 selected lesions (Lesion Analysis Stage 1). A) Bland-Altman plots of median qT1 time in individual WML using different analysis pipelines. B) Bland-Altman plots of individual WML volume using different analysis pipelines. C) Bland-Altman plots of qT1 summary statistics in WML using analysis pipeline 5. See Appendix B.2 for all Bland-Altman values.

small sample size limits conclusions. Lesion volume variability was generally stable across locations, with a few outliers in deep WM and periventricular lesions (Figure 2.11D, bottom). Interestingly, these outliers were categorized as *diffuse*, which tended to have more variability in volume than defined lesions (Figure 2.11E, bottom). In contrast, qT1 variability was highest in lesions with well-defined outlines (Figure 2.11E, top), suggesting that small changes in segmentation could significantly impact qT1 variability.

In Figure 2.12, the analysis was extended to all WML in the test-retest cohort (*Lesion Analysis Stage 2*), simulating the application of these pipelines in large longitudinal datasets. The trends observed in Figure 2.10 largely hold: pipeline 3 exhibits the highest variability in both median qT1 (Figure 2.12A) and lesion volume (Figure 2.12B), while pipeline 2 maintains the most stable median qT1 values. Pipeline 5 ranks second for qT1 reliability and is by far the most consistent in measuring lesion volume.

Importantly, the discrepancy in WML volume test-retest variability for pipeline 4 in Figure 2.12B can be attributed to a single outlier (a very large confluent lesion). Removing this lesion in Figure 2.12B narrows the LoA to (87.9, 80.6), which is much more comparable to Pipeline 5 (60.4, −60.6). Lastly, test-

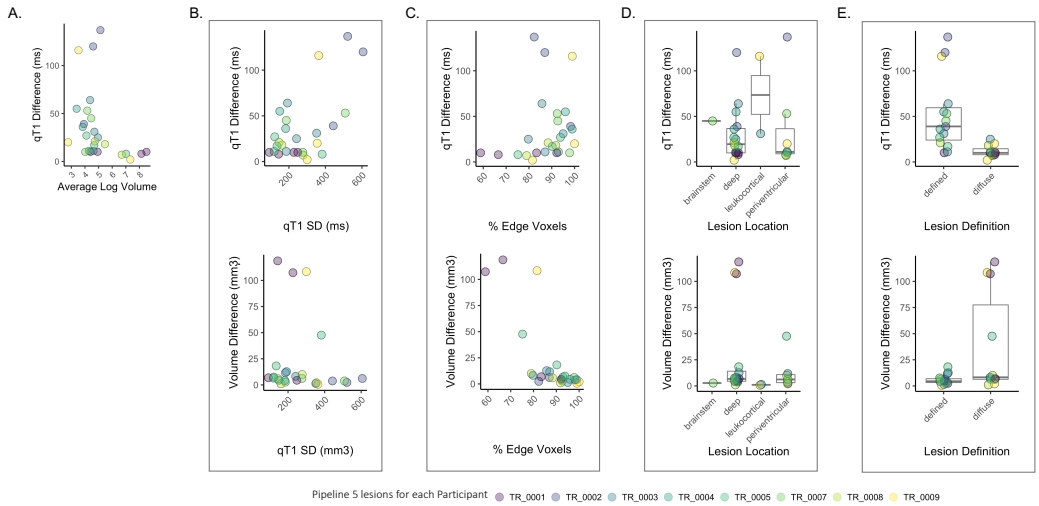


Figure 2.11: Exploratory evaluation of WML characteristics and their association with test-retest variability in median qT1 and volume in 27 selected WML (*Lesion Analysis Stage 1*). *qT1 Difference* and *Volume Difference* are calculated as the difference in median WML qT1 and WML volume, respectively, between session 1 and 2. The average difference refers to the average value between session 1 and session 2 WML. A) Scatter plot of median WML qT1 difference versus average lesion volume. B) Scatter plot of the qT1 (top) and volume (bottom) difference versus the standard deviation of qT1 intensity within the lesion. C) Scatter plot of the qT1 (top) and volume (bottom) difference versus the percentage of edge voxels. D) Boxplots of qT1 (top) and volume (bottom) difference in WML across different locations. E) Boxplots of qT1 (top) and volume (bottom) difference in defined and diffuse lesions. Each point represents a lesion from a participant (color-coded) in the test-retest cohort ($n = 8$) scanned with whole-brain 7T MP2RAGE scans.

retest variability was lowest for qT1 median, 25$^{\text{th}}$ percentile, and mean qT1, with the 25$^{\text{th}}$ percentile showing slightly lower variability, as observed in Figure 2.10.

When looking at all WML, no significant correlations were found between the difference in median qT1 between sessions and average WML volume, qT1 standard deviation, or percentage of edge voxels (Figure 2.12D, E, H). However, there was a positive correlation between the difference in WML volume between sessions and the average qT1 standard deviation in WML (Figure 2.12F, $\rho(412)= 0.1$, p $= 0.03$). A slightly stronger positive correlation was seen between the difference in WML volume and percent edge voxels (Figure 2.12H, $\rho(412)= 0.2$, p $= 0.00001$).

2.2.4 Discussion

In this study, I aimed to evaluate the reproducibility of qT1 measurements using the 7T MP2RAGE sequence in MS participants, focusing on the impact of different image analysis pipelines on qT1 values and lesion volume measurements. GM and WM demonstrated high reproducibility of qT1 measurements compared to CSF and lesions. Image analysis pipelines influenced test-retest

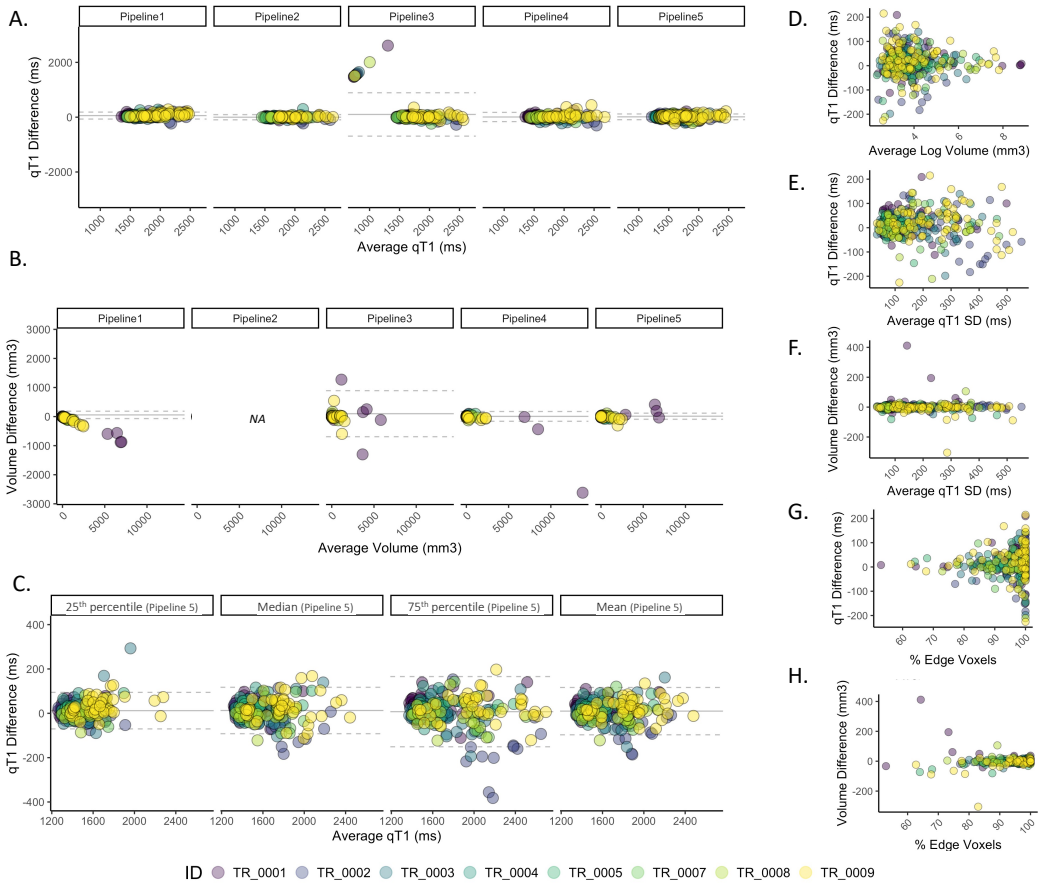


Figure 2.12: Within-scanner Test-Retest Variability of MP2RAGE qT1 times and volume in all WML (*Lesion Analysis Stage 2*). A) Bland-Altman plots of WML median qT1 time using different analysis pipelines. B) Bland-Altman plots of WML volume using different analysis pipelines. C) Bland-Altman plots of qT1 summary statistics in WML using analysis pipeline 5. Pipeline 5 is used for subplots D-H: D) Scatterplot of median qT1 difference between session 1 and 2 vs average log volumes of session in WML; E) Scatterplot of median qT1 difference between session 1 and 2 vs average standard deviation of median qT1 in WML F) Scatterplot of volume difference between session 1 and 2 vs average standard deviation of median qT1 in WML G) Scatterplot of volume difference between session 1 and 2 vs percent edge voxels WML H) Scatterplot of median qT1 difference between session 1 and 2 vs percent edge voxels WML. See Appendix B.2 for all Bland-Altman values.

variability, particularly for qT1 and volume measurements in WML. In pipeline 2, the same segmentation mask was applied to both sessions, and the ROIs were defined by registering each session's segmentation to a halfway space, using only the intersection of the masks to create the final ROI mask. This pipeline, where the segmentation was consistent across sessions, showed the most stable qT1 times across all tissue types. Among the various summary statistics tested, the qT1 median and 25$^{\text{th}}$ percentile were identified as the most reliable indicators for qT1 values across WM, GM, and WML. Pipeline 5 is an automatic lesion boundary correction approach in which lesions from session 1 and session 2 (via registration) are dilated into NAWM and then thresholded using the intersec-

tion point of a bimodal Gaussian mixture model to refine lesion borders. The study demonstrated that this automated lesion segmentation pipeline could achieve test-retest variability in qT1 times and lesion volumes comparable to, or even better than, the gold standard manual segmentation (pipeline 4), suggesting that automated methods are viable alternatives for longitudinal studies in MS.

Among the pipelines evaluated, pipeline 2 exhibited the most stable results across all tissues and lesions, demonstrating relatively low bias and narrow LoA. This stability is likely attributed to the conservative approach in defining the ROI mask. While pipeline 2 was effective in providing stable results, several considerations should be made when applying it to longitudinal studies. Pipeline 2 uses a single segmentation for all sessions, effectively determining qT1 changes in a defined area. However, this does not account for changes in tissue and lesion volume, limiting its suitability in longitudinal studies where tissue regions are expected to undergo changes, such as atrophy or lesion growth. In the context of MS, where tissue dynamics are complex, tissue volume and shape changes are important to capture. For example, MS lesions may expand or shrink over time, which can affect surrounding brain structures. Biological processes like inflammation and oedema can cause voxels to expand or collapse, leading to changes in the shape and location of tissue boundaries, especially at the edges of regions of interest. Using a single mask that represents the union of all time points may constrain the analysis, particularly at biologically significant lesion boundaries.

Additionally, this analysis pipeline may not be ideal for datasets with multiple follow-ups. In the case of two or more MRI follow-ups, the middle time points would be registered to two halfway spaces, while the first and last time points would only be registered to one halfway space. This will bias the results as the middle time points would be weighted more in the analysis. Alternatively, a patient-specific template from all MRI time points could be created. However, participants are not always scanned regularly. Thus, a weighting would need to be included to correct for the varying lengths of time between scans when creating a patient-specific template. If applied to a dataset with ongoing data collection, a new template would need to be re-created each time a new image is added to the dataset, introducing additional complexity to analyse the data continuously.

GM and WM exhibited good reproducibility for qT1 measurements, with qT1 median and 25th percentile values showing stability across all pipelines, maintaining less than ± 31 ms variability (1.3%). Moreover, in the context of MS, qT1 values in GM and WM are more clinically relevant, as they are closely associated with tissue properties such as myelination, axonal integrity, and neuroinflammation. The few studies that have specifically investigated the intra-scanner reproducibility of 7T MP2RAGE qT1 values demonstrated similar high reproducibility of qT1 measurements of GM, WM, and subcortical volumes. In two healthy adults, (Voelker et al., 2016) reported a maximum variability of 0.6% in qT1 values across GM, WM, and subcortical structures, and (Voelker et al., 2021) found that the maximum intra-site repetition errors for subcortical volumes ranged from 0.3% to 3%. Despite those studies focusing on healthy adults, our findings show similar high qT1 test-retest reproducibility in NAWM and NAGM. (Voelker et al., 2021) also found that supratentorial areas exhibited higher intra/inter-reliability than infratentorial areas due to signal dropout in the lower brain regions. This signal loss, particularly in the cerebellum, was also evident in our 7T MP2RAGE scans and could potentially improve reliability if focused exclusively on tissue in the supratentorial regions. However, this aspect was not specifically investigated, as most of the longitudinal analyses will focus on supratentorial lesions due to the limited coverage of infratentorial areas in our T2*-GRE images, which are used in subsequent analyses.

As expected, CSF showed much higher variability across all pipelines, as CSF has distinctly different relaxation properties than other tissues (Condon et al., 1987). As described in the original MP2RAGE publication, 7T parameters were tuned assuming CSF longitudinal relaxation is around 3.35 seconds (s), whereas actual values approach 4-5s (Marques et al., 2010). Since the mapping system (lookup table) used in the MP2RAGE sequence was calibrated around the lower value, actual CSF measurements fall into a part of the mapping curve that is non-linear. This results in tiny fluctuations in the measured CSF signal, leading to large errors in the estimated T1 value. Furthermore, even though MP2RAGE is designed to minimize B1 inhomogeneity effects, residual sensitivity remains: per 20% B1 variation, CSF exhibits 4.2% T1 estimation error compared to 2.5% for GM and 3.8% for WM (Marques et al., 2010). Combined with inherently lower SNR in CSF and the sequence's

emphasis on WM/GM contrast, noise amplification disproportionately affects CSF T1 values.

WML had higher qT1 test-retest variability than GM and WM. WML are much smaller than GM and WM, making them more susceptible to fluctuations in qT1 measurements and variations in tissue segmentation. While no other studies have looked at the test-retest reliability of qT1 in WML, these findings align with previous studies by (Voelker et al., 2016, 2021), which demonstrated that smaller subcortical structures tend to exhibit lower reproducibility for both qT1 times and volumes compared to larger structures, such as cortical areas or larger subcortical regions. Smaller ROIs, like lesions, are more vulnerable to outliers caused by image processing errors or true signal variability. While qualitative results suggested that qT1 standard deviation and percent edge voxels in lesions may influence qT1 variability in a subset of WML (Figure 2.10), Spearman correlations across all WML did not reveal significant associations. However, a slight positive correlation was observed between qT1 standard deviation and WML volume variability, as well as between percentage of edge voxels and lesion volume variability (Figure 2.12). These findings suggest that lesions with a wide range of qT1 values, such as confluent lesions or lesions with more edge voxels, which are typically smaller and more affected by boundary artifacts, may present challenges for consistent volume segmentation. Additionally, periventricular lesions may also affect qT1 variability due to partial volume averaging with CSF, which could skew qT1 statistics to higher values.

Based on WML test-retest findings, some practical considerations may help reduce qT1 and volume variability. Removing small lesions, particularly those composed entirely of edge voxels, is a straightforward step. For example, at 0.7mm isotropic resolution, lesions smaller than $9mm^3$ or $20mm^3$ can consist of up to 100% or 90% edge voxels, respectively, under 6-connectivity criteria. Filtering out lesions below these thresholds can help eliminate those most susceptible to partial volume effects and segmentation instability. Moreover, since MS lesions are typically defined as being $\geq$3mm in diameter, corresponding to a minimum volume of approximately $14mm^3$ at this resolution, a lower threshold between 14 and $20mm^3$ is pathologically meaningful and methodologically robust. Variability of confluent lesion measures could be addressed by including a sensitivity analysis excluding confluent lesions, analysing them separately, or with covariate adjustment. Additionally, removing WML voxels bordering the ventricles can help minimize CSF-related partial volume effects.

WML were analysed in two stages. First, a subset of 26 WML were selected to compare the test-retest variability of qT1 and volume across different pipelines. This approach ensured a fair comparison by using the same lesions across all pipelines, maintaining consistency in lesion selection. Next, the pipelines were applied to all WML to simulate how they would perform in a longitudinal dataset without controlling for key differences in the pipelines, such as how the lesions are spatially split. When comparing the results of these two analyses, some trends were similar in both analyses. Pipeline 5 showed the second-highest qT1 reproducibility after pipeline 2, while pipeline 3 had the most variability in qT1 reproducibility and volume reproducibility in both analyses. In pipeline 1, where individual lesion masks are registered from session 1 to session 2, differences in lesion volume are almost certainly attributable to interpolation errors. While this resulted in low volume variability in *Lesion Analysis Stage 1*, *Lesion Analysis Stage 2* showed increased absolute variability, particularly in larger lesions, likely due to greater cumulative interpolation error with increasing lesion size. Pipeline 4, the pipeline considered a "gold standard" practice of analysis of WML in MS, had much higher variability when applied to all lesions. However, this was due to an outlier in pipeline 4, which caused a large average volume variability. When this large confluent lesion outlier was excluded from the analysis, the reliability was much closer to pipeline 5. This suggests that an automatic lesion analysis pipeline can produce reproducibility similar to that achieved by manual WML segmentation.

A key consideration when applying these lesion segmentation pipelines to large datasets is that they may not always identify the same number of lesions. Lesion splitting, the process of dividing the total WML mask into individual clusters, is a crucial and difficult step when comparing lesions between sessions. In all pipelines, FSL's *cluster* tool is used to split the total WML mask into clusters based on connectivity. In pipelines 1 and 5, the WML mask is split in session 1 space, and individual lesions are then registered to session 2. This approach allows lesions found in session 1 to be analysed across both sessions, but does not account for new lesions in subsequent sessions. However, the approach keeps track of individual lesions across time points. In contrast, pipelines 2, 3, and 4 process session 1 and session 2 data separately, splitting the total lesion masks into clusters and matching lesions based on overlap. This can lead to different numbers of lesions identified in each session.

Additionally, if the lesion in one session is split into multiple lesions in the second session, this can result in incorrect labelling. Registration mismatches can also misalign the sessions, affecting lesion matching. In pipelines 3 and 4, this could explain some of the high variability in WML qT1 and volume. Pipeline 2, however, only includes the intersection of the session 1 and session 2 masks, potentially lesions, particularly smaller lesions. While this pipeline had the highest qT1 reproducibility, it also had the fewest lesions included in the analysis in Figure 2.10. These challenges highlight important limitations when applying these pipelines to large datasets.

In GM, the qT1 median demonstrated the best test-retest variability, likely due to its robustness against outliers in either direction, making it more stable across repeated measurements. For WML, both the 25^{th} percentile and qT1 median had the highest reproducibility, likely because these statistics reduce sensitivity to partial volume effects with areas of high signal, like CSF; partial volume can significantly impact measures like the mean or 75^{th} percentile, especially when lesions are located near the CSF or at tissue boundaries. These findings suggest that the qT1 median and 25^{th} percentile are the most reliable summary statistics for capturing qT1 variability in MS lesions. Furthermore, the 25^{th} percentile has been shown to have the best statistical sensitivity for detecting changes in lesions (Reich et al., 2015), likely because it captures the lower tail of the signal distribution, which corresponds to tissue at the lesion edge where biologically meaningful changes often occur.

In this section, the test-retest variability of qT1 metrics in GM, WM, CSF, and individual lesions across various analysis pipelines was evaluated. While our primary clinical questions are related to individual lesions over time, qT1 in cortical tissues are also of interest, as disease-related changes may affect these tissues and provide valuable insights into their relationship with MS lesions. For lesion analysis, each pipeline offers distinct advantages and limitations. For the purposes of this thesis, pipeline 5 was selected for the subsequent analyses based on several key considerations. First, I prioritized the need for an automated pipeline capable of scaling to large datasets. Second, my analysis aims to focus on longitudinal data, where images are collected at multiple, non-uniform timepoints, with new data continuously being added. As noted earlier, pipeline 2 would be less suitable for this due to the complexity of introducing template spaces. Third, I sought an approach that could track both lesion volume and qT1 changes, which was not feasible with pipelines 1 and 2. Although Pipeline

5 does not detect new lesions, this limitation is not relevant for my research interests of tracking changes in chronic lesions, where paramagnetic rims may have formed. Importantly, variability in qT1 measurements of up to 100 ms may be attributable to noise, particularly in smaller lesions. These findings are important in interpreting clinically relevant changes in disease progression and treatment response.

2.3 Development of a Visualization Application to Streamline Multi-modal Longitudinal Lesion Assessment

2.3.1 Introduction

The ability to visually assess and quantify MS lesions through MRI is critical for both clinical and research purposes. Radiologically determined lesion characteristics, such as PRLs, are closely associated with worsening clinical symptoms (Absinta et al., 2016; Thaler et al., 2017). The PRL can be visualized around certain lesions due to iron-laden macrophages aggregating at the lesion edge, indicating ongoing chronic inflammation. PRLs are visible using MRI techniques like T2*-GRE phase, with the rim gradually diminishing over time, disappearing with a half-life of about seven years (Absinta et al., 2021). Thus, the disappearance or reduction in conspicuity of the paramagnetic rim around lesions may be a potential endpoint in clinical trials targeting chronic inflammation.

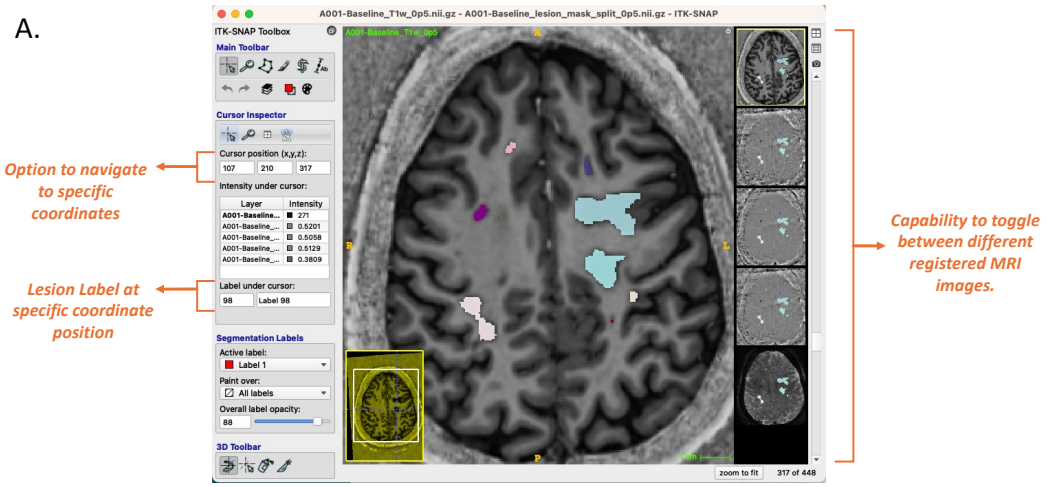
Automated methods for detecting PRLs have been developed, mainly for 3T MRI using T2*-GRE phase and QSM techniques. However, challenges such as limited annotated datasets and substantial inter-rater variability in PRL ratings continue to hinder widespread use of automated PRL detection (Barquero et al., 2020; Zhang et al., 2022). As a result, expert visualization remains the standard, though it is a labour-intensive and time-consuming process (Kolb et al., 2022).

Despite their importance, visual ratings of lesions present several challenges. Raters must navigate large, complex MRI datasets, and sharing datasets can be complicated due to identifiable features in whole-brain MRIs, even when metadata is anonymized. The process of assessing longitudinal changes in PRLs typically involves downloading large datasets and performing image processing

steps, including normalization and registration. Various visualization software tools, such as *FSLeyes*, *MIPAV*, *ITK-SNAP*, and *AFNI*, aid in the analysis of MRI data by enabling the comparison of lesions across modalities and time-points, either through image registration or by viewing already registered images.

An example of an approach to analyse longitudinal PRL data is shown in Figure 2.13. Before loading MRI images into an MRI viewer, images are registered within and between sessions for varying modalities and longitudinal data. Then, all images can be loaded into an MRI viewer along with a lesion mask with different labels for each lesion (Figure 2.13A). A spreadsheet is created with centre of gravity lesion coordinates for each lesion and shared with the neurologists and radiologists, who can add ratings and notes to the spreadsheet (Figure 2.13B). However, this process is time-consuming, prone to data entry errors, and demands meticulous attention to detail to ensure the accurate identification of all lesions.

The purpose of this section is to develop a web-based visualization application to streamline multimodal longitudinal lesion ratings. By combining advanced imaging analysis with a structured data management system, the application aims to address a critical gap in the existing MRI visualization tools. The application development involved close collaboration with neurologists who provided input and feedback to ensure that the tool would include key features and address unmet needs of the current approach.



Baseline PRL Ratings

Participant	Lesion	Lesion Center of Gravity (Voxels)			PRL (0 or 1)	Comments
A001	1	211.3	167.6	297.0	0	
A001	2	226.6	271.0	274.9	1	
A001	3	106.7	308.4	287.6	1	subtle
A001	4	223.0	205.9	293.1		
A001	5	199.2	224.6	308.8		
A001	6	186.0	295.0	299.9		
A001	7	180.9	212.9	295.0		
A001	8	46.8	228.0	291.9		
A001	9	123.8	238.5	269.0		
A001	10	86.2	241.0	216.0		

Figure 2.13: Example of a current method for visually rating lesions. A) Registered MRI images can be loaded and toggled in MRI visualization software (e.g., ITKSnap). A lesion mask is used overlayed to identify lesions based on labels, and coordinates are entered to navigate to specific locations. B) A spreadsheet is provided to raters, with columns for identifying the lesions on the left and columns for visualization ratings and comments on the right.

2.3.2 Methods

2.3.2.1 Overview of Application Design and Use

I designed a web-based visualization application using Flask, a lightweight and open-source web framework based on Python (Flask version 2.3.2, Pallets, https://github.com/orgs/pallets/repositories). Flask was selected for its simplicity, ease of use, and the ability to create an application without significant web development experience. The application enables users to view and rate MRI lesions through a dynamic and interactive web interface. Figure 2.14 provides an overview of an example webpage, highlighting key features of the visualization application. The application displays lesion data in a paginated format, with each page presenting a set of lesions to be rated. A dynamic form is generated for each page, allowing users to select lesion characteristics (such as changes in the paramagnetic rim) and provide comments. Additionally,

the application supports multiple raters, who can identify themselves using a drop-down menu to select their name.

The visualization application was designed with neurologists to incorporate key features for lesion rating. One design requirement is the ability to include multiple timepoints and contrasts for each lesion. Different MRI contrasts may be needed for assessing different lesion characteristics, and multiple timepoints allow for longitudinal ratings. The application's design aggregates the contrast and time point data in a way that allows easy viewing: timepoints are represented along the columns, and MRI modalities are placed on different rows. Brightness and contrast sliders enable raters to adjust settings, improving their ability to visualize the lesion and make accurate ratings. The clickable PNG images allow users to scroll through axial PNG slices. This ensures raters can view the lesion comprehensively and with the necessary contrast to make informed ratings.

Another important design consideration is anonymization, particularly for clinical trials. The application can anonymize images both in terms of timepoints and lesions. When creating the PNG images for the app, the timepoints can be appended in any order, and the order will be stored in a database compatible with the Flask app. On the web interface, timepoints are anonymized as *Column 1*, *Column 2*, etc., while the application ensures the order is recorded in the back-end and included in the final rating database. This feature reduces potential bias that might result from the chronological order of timepoints.

Lesion anonymization is another critical aspect of the design. To reduce bias, lesions are randomized between participants. Each page may display lesions from different participants, and the back-end matches the lesion ID with the participant ID when submitting ratings. This randomization is particularly important in clinical trials, where it allows for the comparison of lesions from different groups, such as drug versus control participants. The ability to strip away participant identifiers from individual lesions helps avoid bias, especially when evaluating lesions in isolation without the full brain context. Additionally, the design facilitates intra-rater reliability ratings by allowing lesions to appear multiple times without revealing their original identifiers.

The application also supports multiple raters by providing a drop-down menu where raters can identify themselves. This feature is essential for studies that require the input of several clinicians or experts. Additionally, to accommodate raters who need to come back to their ratings after a session, the

application divides the dataset into smaller, digestible pages, each containing six lesions. This allows neurologists and radiologists to complete ratings in multiple sessions without feeling overwhelmed by the entire dataset.

Finally, the application ensures raters can track their progress. For each page, once a rating is submitted, the name and time of submission are displayed at the top of the page. This functionality helps raters keep track of what they have already rated and prevents them from revisiting or altering their previous responses.

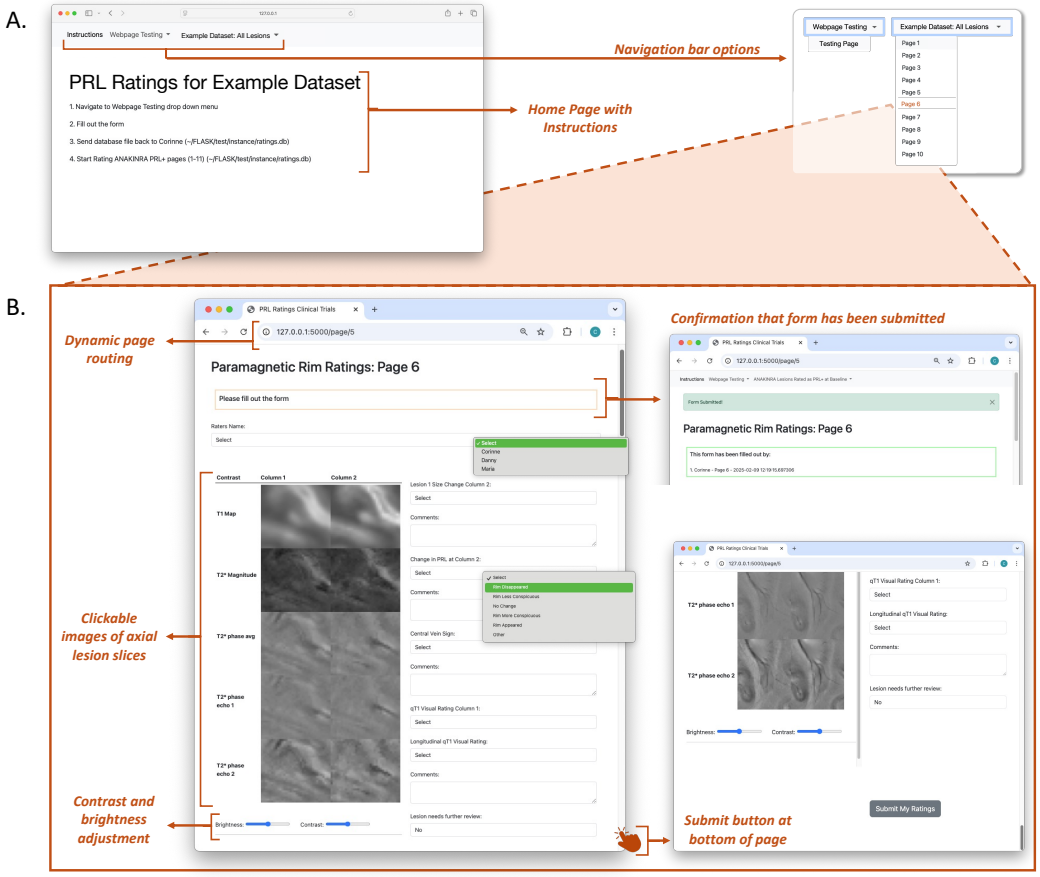


Figure 2.14: Overview of Flask Visualization Web Page Application. A) The homepage displays rating instructions with a navigation bar offering three options: Instructions (displays the homepage), Webpage Testing (for testing the form), and Example Dataset: All Lesion Rating Pages (with a drop-down menu for 10 pages). B) Example Rating Page showing the first lesion. Dynamic routing uses the rating page number as the URL. Axial slices of lesions with different MRI modalities (rows) and timepoints (columns) are displayed. The first image is the lesion's midpoint, and users can click through axial slices. Adjustable sliders for brightness and contrast are provided. On the right of the image, a corresponding form is provided drop-down menus for ratings and text boxes for comments. The submit button is at the bottom of the page. When all forms are filled and the submit button is clicked, a success message is displayed. If ratings are not submitted, an orange box prompts users to submit; once submitted, it turns green, and the submission is logged with the rater's name and timestamp.

2.3.2.2 Application Development

MRI Image Processing: MRI images need to be processed to prepare the data for display in the Flask application (Figure 2.15A). While the specific preprocessing steps may vary depending on the dataset, they generally consist of four main processes: image preprocessing, registration, intensity normalization, and lesion segmentation.

Image preprocessing typically involves bias field correction and steps to facilitate registration, such as neck cropping and skull stripping. All images need to be registered into a common subject space, which includes within-session registration for datasets with multiple modalities and between-session registration for longitudinal datasets. All images need to be resampled to a consistent voxel size to ensure comparable ROI across modalities and timepoints. Intensity normalization needs to be performed to standardize intensity scales across images, ensuring all relevant features are visible regardless of modality. Finally, a lesion mask for each participant is needed with a unique label for each lesion, the labels not needing to follow a consecutive numbering scheme.

Generation of Flask-Compatible Visual Assets: After the MRI dataset is preprocessed, it is converted into Flask-compatible visual assets (Figure 2.15B). To automate this conversion, I developed a Bash script that processes MRI data into the appropriate formats and file structure for visualization within the Flask application. The script is designed to handle data for a single participant from the dataset. As illustrated in Figure 2.15A, the script takes six inputs: (1) the folder path containing the preprocessed MRI data, (2) the folder name for the given participant, (3) the file name for the participant-level lesion mask, where each lesion has a unique label, (4) a space-separated list of folder names for the timepoints, (5) a space-separated list of MRI contrasts, as well as (6) the output directory path where the results will be saved.

Although the folder and file names can be customized, the required folder structure for the input data is outlined in Figure 2.15B. Additionally, the script can accommodate any number of timepoints and contrasts. The output includes the necessary files and structure that are compatible with the Flask application, requiring only that they be moved to the *Static* folder (Figure 2.15) for use in the web interface.

Briefly, the Bash script iterates through each unique lesion ID in the lesion mask, performing a series of steps for each lesion label. First, it uses FSL's

fslstats tool to identify the smallest bounding box encompassing non-zero voxels. The script then extracts the lesion's ROIs by applying the coordinates of the bounding box, with additional padding, to all MRI images using FSL's *fslroi* tool. For all ROIs, it generates PNG images for each slice using FSL's *slicer* tool and resizes the resulting PNG with the Scriptable Image Processing System (SIPS) to ensure consistent image dimensions. The generated PNG images for each timepoint and contrast are then combined into a single composite image, producing composite PNGs representing the lesion's axial slices.

Additionally, the script creates a participant-level text file containing a list of dictionaries in a Python-compatible format. As the script processes each lesion, it adds a dictionary to this text file, including the subject ID (provided via the -P input), the lesion label, the file path to an image representing the lesion's midpoint, and a list of file paths corresponding to all axial PNG images for that lesion.

Anatomy of the Flask App: The Flask application is structured with various files that must be organized in a specific format (Figure 2.15C). The *app.py* file is the core of the application, handling the initialization of Flask, routing, and connecting the back-end with the front-end. The application stores lesion ratings in an SQLite database, dynamically displays lesion data and forms through HTML pages, and uses Flask forms and Jinja2 templates to render content. Routing allows users to navigate between different lesion pages, with form submissions being saved in the database.

The application uses an SQLite database, which is directly built into Python, to store user ratings for MRI lesions in a table defined in the *app.py* file. This table includes columns for participant and lesion identifiers, as well as user input fields such as lesion ratings and usernames. SQLAlchemy, integrated with Flask, initializes and manages the database, allowing for operations like adding, querying, and updating entries. The database file, *rating.db*, is stored in the instance folder and is updated in real-time as users submit ratings. Upon completion, the database can be easily shared or exported to formats such as CSV or Excel for further analysis.

The web application's form is created using WTForms, a flexible form validation and rendering library in Python, to allow users to input ratings for each lesion. The form is dynamic so that it can adapt based on the data provided; this allows it to be flexible for different datasets with varying amounts of lesions. The form has various classes for different ratings, including drop-down

options for lesion ratings and text area fields for comments. When the form is submitted, the data is validated and added to the SQL database as new entries. The form is dynamically populated based on the number of lesions, making it flexible for varying datasets.

The HTML files, located in the template folder, define the layout and structure of the Flask app's web pages using the Jinja2 templating engine to dynamically generate content. The application only needs four HTML files. The *base.html* file is a template that provides a common structure (header, navigation, footer) for all web pages. The *index.html* file contains the homepage content, displaying instructions for navigating the PRL Ratings form. The *navbar.html* creates a responsive navigation bar with links to different sections, utilizing Bootstrap for design components. Finally, the *template_participant.html* renders the lesion rating page, where users can interact with images (adjust brightness/contrast, click through slices) and submit ratings through dynamically generated forms. The flash messages that notify users upon successful form submission are also coded here.

Axial lesion PNG images and corresponding participant-level text files are stored in the static folder, linking lesion data to image locations. The *app.py* file collects and combines all lesion information from *imagelist.txt* files across the static folders into a large database. After the data is gathered, lesions are randomly shuffled between participants. The shuffled database is then divided into manageable chunks of six lesions, creating paginated "pages" for review. This pagination presents the data in smaller, digestible sections. The final shuffled and structured database is used to dynamically populate *template_participant.html*, where users can rate lesions and submit feedback.

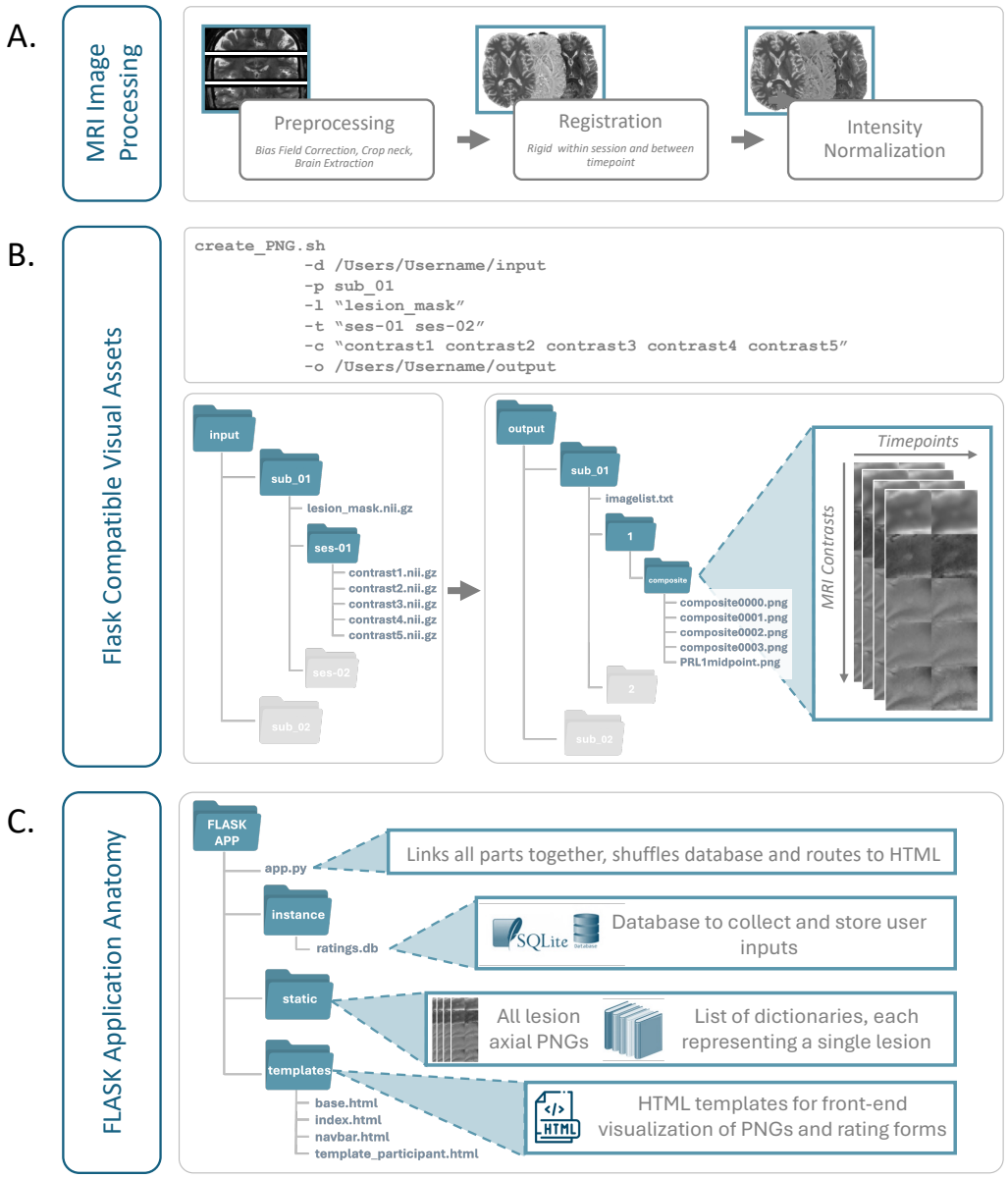


Figure 2.15: Overview of the steps involved in developing the Flask application. A) MRI images are pre-processed to ensure spatial and intensity standardization across all modalities. B) MRI images are converted into Flask-compatible PNGs using a bash script (top). The script requires a standardized input file structure (left) and generates an output folder with PNGs of each lesion's axial slices. C) The final file structure for the Flask application is shown, including the main *app.py* file, the SQL database, static images (generated in B) and HTML templates for the front-end aspects of the webpage.

2.3.2.3 Example Clinical Trial Implementation

This section provides a general overview of how a Flask application was generated for the ATaC-MS clinical trial primary outcome. While Chapter 3.2 outlines the full image processing pipeline and trial-specific insights, a brief overview of the image processing steps relevant to the Flask application is pro-

vided below.

A total of 30 lesions from 5 participants were rated across two timepoints. For each timepoint, five contrasts were available: a qT1 map, a T2*-GRE magnitude image (averaged across echoes), and the corresponding unwrapped T2*-GRE phase images (from two individual echoes and an echo-averaged map). At baseline, each participant had a PRL+ lesion segmentation mask, with each lesion assigned a unique label.

The MRI preprocessing steps are shown in Figure 2.15A. T2*-GRE magnitude and phase images, which were acquired in slabs, were bias field corrected and stitched together. T1 maps were cropped to remove the neck region and then up-sampled to a higher resolution to preserve the high in-plane resolution of the T2*-GRE images. Follow-up T1 map scans were rigidly registered to the baseline scan's T1 map. T2*-GRE magnitude images were rigidly registered within the same session to the T1 map and then longitudinally registered to the baseline T1 map by combining and applying both within- and between-session registration matrices. Intensities were normalized to a range of 0 to 1 (see Section 3.2.2.3 for normalization details).

To generate PNG images for visualization, the script shown in Figure 2.15B was run twice: once with timepoints input in the correct chronological order (Baseline, Follow-up) and once with the timepoints reversed (Follow-up, Baseline). This approach ensured that each lesion was shown twice in the Flask app, rated in both chronological and reverse chronological orders to minimize bias in the ratings. All lesions were pooled and randomly scrambled in the *app.py* file to further prevent rating bias. Two raters rated all lesions and provided qualitative feedback on the utility of the app.

2.3.3 Evaluation and Feedback

I developed a web-based visualization application to streamline multimodal longitudinal lesion ratings, specifically for PRLs. To evaluate its effectiveness, I collected qualitative feedback from neurologists after they used the application to assess primary outcomes in two clinical trials described in Chapter 3.4.3. Key features that stood out were the ease and flexibility of anonymization and randomization, both for lesions between participants and for timepoints. Both raters reported that the application allowed for easier ratings, enabling more assessments in a shorter amount of time. One of the primary reasons for this efficiency, compared to the process shown in Figure 2.1, was that users did not

need to spend time setting up the viewer, searching for lesions, or arranging different windows for contrasts and timepoints.

Additionally, the app's file size is approximately significantly smaller than downloading the MRI datasets used in our clinical trials. For example, the folder containing all MRI images for the primary outcomes dataset of the ATaC-MS clinical trial is 10.15GB (5 participants with 2 timepoints) while the Flask application takes up only 14MB. Scaling this to larger longitudinal datasets with over 80 participants (about 752GB), the Flask application would only take up 1.1GB, a 680x reduction in size.

Importantly, feedback from raters highlighted that the quality and sensitivity of ratings depend heavily on the quality of image registration. As discussed in Chapter 2.2, inter-modality registration at 7T is particularly challenging due to field-induced distortions. Rigid registration was used for both within-modality and between-session registration. Aligning T2*-GRE magnitude images with qT1 maps was difficult as some regions aligned well, while others remained misaligned due to distortion. Distortion-related registration issues can sometimes be mitigated by increasing the degrees of freedom or using non-linear registration. For example, in Chapter 2.2, combining affine registration with a boundary-corrected pipeline (pipeline 5) ensured that follow-up lesion masks accurately captured lesion boundaries (although in that context, the registered image itself was not the output of interest). For multimodal longitudinal visual ratings using the clinical trial datasets, rigid registration was more suitable for several reasons. First, some T2*-GRE image slabs were misaligned during acquisition, introducing gaps that could confound non-rigid registration. Second, in addition to PRL status, lesion size changes were part of the ratings. Non-rigid transformations risk introducing artificial warping of lesions (potentially more than the original geometric distortions) which could bias lesion size assessments. Despite the imperfections in registration, in the ATaC-MS dataset, only 1 out of 60 cases was considered un-ratable because of mis-registration. Both raters agreed that all lesions rated as having a change in PRL status need to be validated with the original scans. While this required us to share the original MRI files, it was only a few cases. Detecting changes in PRL is challenging and often very subtle, verifying changes using the MRI scan in other lesion markers may not be needed.

Another challenge encountered by the raters was setting up the installation to be able to use the Flask application, even with detailed instructions. Both

raters needed to download and set up an Anaconda environment with all dependencies. However, the installation only needed to happen once. Hosting the application on a shared server would make it more accessible and avoid the need for raters to deal with the installation on their local computer. Furthermore, the application could be hosted online, as the lesion PNG files are anonymized, scrambled and have no identifiable information of participants. This could facilitate widespread access to colleagues in different institutions and industries facilitating collaboration among raters and enable the aggregation of data for larger studies. To host the webpage online, additional security measures would need to be added. A login page that allows different users to log in and access the protected page could be set up.

In summary, the development of this Flask application for visual rating of MS lesions represents a significant advancement in addressing the challenges of manual lesion assessment. By providing a user-friendly interface, the application enhances ease of lesion ratings across multiple raters, contributing to better clinical research design.

Chapter 3

Longitudinal Assessment of Quantitative T1 and Volume in Chronic White Matter Lesions at 7T

Contents

3.1 Longitudinal Changes of Quantitative T1 and Volume in Chronic White Matter Lesions in a 7T Natural History Dataset 79

3.2 Effect of Anakinra on Chronically Inflamed White Matter Lesions in Multiple Sclerosis 101

3.3 Effect of Tolebrutinib on Chronically Inflamed White Matter Lesions in Multiple Sclerosis 124

3.4 Discussion of MRI Results from Clinical Trials . . 145

This chapter investigates the longitudinal evolution of quantitative T1 (qT1) values and lesion volume in chronic white matter lesions (WML) in multiple sclerosis (MS). A subset of chronic lesions exhibits chronic inflammation driven by activated microglia and macrophages (Dal-Bianco et al., 2017), which can be visualized on susceptibility-sensitive magnetic resonance imaging (MRI) as a paramagnetic rim. Lesions with a paramagnetic rim (PRL or PRL+) are associated with greater disability and tissue damage compared to chronic lesions without a visible rim (PRL-) (Absinta et al., 2016; Dal-Bianco et al., 2021,a; Toru Asahina et al., 2024). Chronic inflammation is thought to contribute to progression independent of relapse activity (PIRA), a major driver of long-term disability that is not addressed by currently approved disease-modifying therapies (DMTs). Despite their clinical significance, the long-term trajectories

of PRL+ and PRL- lesions remain poorly characterized. Characterizing the trajectories of PRL+ and PRL- lesions is important to determine the extent to which these lesions persist and evolve, to reveal mechanisms of gradual tissue damage or repair, and to inform therapeutic strategies aimed at preventing further injury and supporting repair (Geladaris et al., 2022; Piehl, 2021). Currently, no treatment effectively targets chronically inflamed lesions in MS, even though they persist over time and contribute to disease progression. Given the role of activated macrophages and microglia in sustaining this inflammation, their short-term depletion or modulation may prevent further tissue damage and potentially support endogenous repair (Geladaris et al., 2022; Piehl, 2021). Characterizing the long-term trajectories is critical to assess whether PRLs represent a suitable endpoint for clinical trials and to provide a benchmark against which the effects of emerging therapies targeting chronic inflammation can be evaluated.

The primary aims of this chapter are to characterize how PRL+ and PRL- evolve and to evaluate whether investigational treatments can modulate the paramagnetic rim, PRL+ qT1 and volume. Using image analysis tools and pipelines developed in Chapter 2.3.3, in this chapter I describe the assessment of chronic WML evolution within a natural history cohort and in two clinical trials targeting chronically inflamed lesions. Analyses across the three studies were conducted in parallel, using similar but slightly adapted image analysis pipelines to accommodate study-specific differences. Statistical analyses for the clinical trials were conducted in collaboration with a statistician (T.W.), whose statistical analysis pipelines were also adapted and applied to the natural history study.

In Section 3.1, I analyse longitudinal qT1 changes in PRL+ and PRL- chronic WML using a large natural history dataset. I presented preliminary results from this study at conferences (ACTRIMS 2022, ECTRIMS 2023, and ISMRM 2024).

Sections 3.2 and 3.3 describe the analyses I conducted as part of my involvement in clinical trials examining the impact of two investigational DMTs, anakinra and tolebrutinib, on chronic inflammation. Specifically, these analyses assess the ability of either DMT to modulate the paramagnetic rim and their effects on lesion volume and qT1 in PRL+ compared to PRL- lesions. Each section includes a brief overview of the therapeutic agent, details of the MRI-related aspects of the clinical trial design, image processing methods (largely

consistent across trials), and a summary of key findings. Preliminary results from analyses I conducted on the MRI data from these studies have been presented by the principle investigator and lead associate investigators of the clinical trials, on behalf of all the co-authors at ACTRIMS 2024 and 2025.

Finally, in Section 3.4, I provide an integrated discussion of the results of both clinical trials. This section synthesizes the results on qT1 and volume evolution in PRL+ versus PRL- from the interventional studies in relation to the natural history cohort, and reflects on the implications for future therapeutic strategies targeting chronic inflammation in MS.

3.1 Longitudinal Changes of Quantitative T1 and Volume in Chronic White Matter Lesions in a 7T Natural History Dataset

3.1.1 Introduction

PRL+ represents chronic active inflammatory lesions in MS and have been increasingly recognized as clinically relevant markers of disease severity. Cross-sectional and longitudinal studies consistently report that PRLs are associated with greater disability and are predictive of worsening clinical outcomes (Reeves, Weinstock, Zivadinov, Dwyer, Bergsland, Salman, Schweser, Weinstock-Guttman, Benedict and Jakimovski, 2023; Toru Asahina et al., 2024). PRL burden, whether assessed as a binary presence/absence (Calvi et al., 2023), a categorical grouping (e.g., $\geq$ 4 PRLs) (Absinta et al., 2019; Altokhis et al., 2022), or a continuous lesion count (Coffman et al., 2022), is associated with higher expanded disability status scale (EDSS) scores and/or faster EDSS progression. In addition to EDSS, some studies have reported associations between PRL burden and cognitive impairment, including lower scores on the Symbol Digit Modalities Test (SDMT) and the Paced Auditory Serial Addition Task (PASAT) (Absinta et al., 2019). However, some inconsistencies remain (Toru Asahina et al., 2024). For example, Blindenbacher et al. (2020) reported no significant associations between PRL count or presence and performance-based outcomes, including the MS functional composite score (MSFC), 9-hole peg test (9HPT), timed 25-foot walk (25FW), PASAT, or SDMT (Blindenbacher et al., 2020).

Nevertheless, the impact of PRL+ on disability appears to extend beyond relapse-associated progression. One study demonstrated that individuals with a greater number of PRL+ had increased odds of experiencing PIRA (Reeves, Bartnik, Jakimovski, Mohebbi, Bergsland, Salman, Schweser, Wilding, Weinstock-Guttman, Dwyer and Zivadinov, 2024). Another study showed that the resolution of PRL+ was associated with improved clinical outcomes and lower PIRA rates, independent of overall lesion burden or new lesion formation (Reeves, Mohebbi, Wicks, Salman, Bartnik, Jakimovski, Bergsland, Schweser, Weinstock-Guttman, Dwyer and Zivadinov, 2024). These findings suggest that chronic inflammatory activity at the lesion edge, rather than acute focal inflammation, can be a key contributor to progressive clinical decline.

PRL+ lesions exhibit destructive tissue characteristics. Quantitative 7T MRI studies show that PRL+ have prolonged qT1 times compared to PRL- lesions (Absinta et al., 2019; Dal-Bianco et al., 2021; Kolb et al., 2021) and lower fractional anisotropy (FA) and higher radial diffusivity (RD) relative to total lesion burden (Elliott et al., 2023). This evidence of tissue damage, together with PRL+ associations with worsening disability, has prompted increasing efforts to characterize their longitudinal behaviour.

Studies aiming to establish the onset of rim development show that paramagnetic rims tend to emerge from enhancing lesions with a centripetal enhancement pattern, where gadolinium enhancement progresses inward from the lesion edge (Absinta et al., 2016; Al Gburi et al., 2025; Boffa et al., 2025; Wenzel et al., 2022). Enhancing lesions that later become PRL+ are distinguished early on by greater T1 hypointensity, longer qT1 times, lower semi-quantitative T1w/T2w ratios, lower normalized magnetization transfer ratios (nMTR), and larger volumes compared to enhancing lesions that do not evolve into PRLs (Absinta et al., 2016; Al Gburi et al., 2025; Boffa et al., 2025; Clark et al., 2023; Elkady et al., 2025; Wenzel et al., 2022). This supports the idea that PRL+ have a more aggressive inflammatory profile even during the acute phase. Among these features, lesion size appears to be the most predictive early marker; each millimeter increase in diameter has been linked to a 44% higher likelihood of developing a paramagnetic rim (Al Gburi et al., 2025).

While the transformation from enhancing lesions into PRLs is becoming better understood, fewer studies have examined the long-term trajectory of PRLs after their formation. At 3T, studies with less than 2-year follow-up

have reported progressive tissue damage in PRL+ lesions, as reflected by semi-quantitative MRI measures, diffusion measures, and volume changes (Elkady et al., 2025; Elliott et al., 2023; Weber et al., 2022). Two 7T studies with longer follow-up durations, but smaller sample size, have shown similar evolution of chronic lesions. A study of 33 MS patients followed for up to seven years demonstrated that lesion volume trajectories differ by PRL status and disease subtype, with significant interactions between time and rim status observed in secondary progressive MS (SPMS), but not in relapsing-remitting MS (RRMS) (Dal-Bianco et al., 2021). Among SPMS participants, PRL+ and PRL- lesions exhibited distinct volume evolution patterns, with PRL+ showing a non-significant tendency to grow, while PRL- shrank. In the same study, categorical analysis using a 10% expansion threshold to define lesion growth revealed that, in RRMS patients, approximately 50% of PRL+ lesions enlarged while 30-40% shrank. In contrast, 82% of PRL+ lesions grew in SPMS participants, suggesting that even within the PRL+ category, there is heterogeneity in lesion evolution. The lack of observable expansion in some PRL+ may be due to expansion at the lesion edge being offset by concurrent shrinkage at the lesion core, potentially caused by severe tissue damage and structural collapse (Sethi et al., 2017). Another 7T study of 23 participants with a mean follow-up of 14.5 years found that PRL+ lesions grew at an estimated rate of +2.2% per year, whereas PRL- lesions shrank at -3.6% per year (Absinta et al., 2019). Additionally, expanding lesions, regardless of rim status, exhibited higher baseline and longitudinal qT1 values, indicating greater tissue damage (Absinta et al., 2019).

Evidence suggests that PRLs represent a more destructive lesion phenotype characterized by greater and more persistent tissue damage over time. However, current studies are limited either by small participant numbers or by relatively short follow-up periods. Consequently, there remains a critical need for larger-scale, long-term investigations to better understand PRL evolution and their precise contribution to MS progression.

Lesion location may also influence the longitudinal evolution of chronic WML. While PRLs have been reported across various brain regions, their distribution remains incompletely characterized, with studies identifying differing proportions in periventricular (PV), juxtacortical (JC), and deep white matter (DWM) regions (Guo et al., 2021; Kolb et al., 2021; Meaton et al., 2022). In one study, PV PRLs were associated with increased annual relapse rates

while JC and DWM PRLs were not, suggesting a potential link between PRL location and clinical activity (Guo et al., 2021). In another study, chronic WMLs located near the ventricles have been shown to exhibit more destructive pathology, including greater edge expansion and accelerated tissue loss within the lesion core (Klistorner et al., 2022). These findings raise the possibility that PV PRLs may be particularly aggressive, underscoring the need to better understand how lesion location influences PRL pathology and evolution.

Additionally, recent work has emphasized the value of separating lesions into core and rim regions, as rim-based analyses have uncovered biologically relevant differences not captured by whole-lesion statistics (Elkady et al., 2025). Given that demyelination and repair processes may preferentially occur at the lesion edge and the core is more likely to reflect irreversible axonal loss (Klistorner et al., 2021), longitudinal assessment of volume and qT1 within these subregions may offer deeper insights into chronic WML.

In this section, I compare chronic PRL+ and PRL- qT1 and volume on 7T images both cross-sectionally and longitudinally, to understand how these lesions evolve over time and contribute to progressive tissue damage in MS. In a large sample of participants with long-term follow-up. Additionally, I explore whether these patterns vary by lesion location and edge/core substructures.

3.1.2 Methods

3.1.2.1 Natural History Dataset Curation

A retrospective longitudinal dataset was curated from the Translational Neuroradiology Section's (TNS) natural history MRI database (Appendix A.1) by systematically reviewing over 19 terabytes of MRI data acquired before June 2023. Metadata from all MRI scans were extracted and filtered in *R* to identify longitudinal 7T brain MRI in participants with a recorded MS diagnosis. Participants were included if they had at least two 7T Magnetization Prepared 2 Rapid Acquisition Gradient Echo (MP2RAGE) scans ($0.7\,\mathrm{mm}^3$ isotropic resolution, without gadolinium injection and excluding developmental or non-standard variants) acquired at least 6 months apart, and a gradient recalled echo T2*-weighted (T2*-GRE) baseline scan that covers the supratentorial brain either via whole-brain acquisitions or with a minimum of three slabs spanning the full supratentorial region (see Appendix A.3 - A.3.1 for final dataset details). A minimum follow-up interval of 6 months was chosen to

increase the likelihood of capturing the slow changes characteristic of chronic lesion evolution. All MRI scans underwent a quality control step in which PNG images were generated for a single central slice in each plane (sagittal, axial, and coronal) as well as for a series of axial slices sampled every few slices. Generating PNGs instead of downloading the full dataset allowed visual quality control while minimizing data size. Scans with significant artefacts or motion were excluded.

When available, clinical history data were extracted for each participant, including the participant's approximate age at MS onset, DMT history, and diagnostic subtype. Each participant was assigned a treatment category corresponding to the DMT they were on for the majority of the MRI follow-up period. The treatment categories were based on annualized relapse rate (ARR) reduction, following Singer et al. (2024): untreated, low efficacy, moderate efficacy, and high efficacy. In cases where treatment was evenly split between two categories, the most recent treatment was used to assign the participant-wise category. Given the complexity and variability of individual treatment histories, where switching treatments or experiencing treatment interruptions is not uncommon, this simplified categorization created a manageable control variable for statistical modelling. MS diagnostic subtype (RRMS, PPMS, or SPMS) was confirmed using clinical notes. Changes in the recorded diagnosis subtype over the course of MRI follow-up were documented.

3.1.2.2 MRI Image Processing

Figure 3.1 provides an overview of the MRI image processing steps. All MRI scans underwent bias field correction, except for the qT1 map, which is already bias field corrected (Marques et al., 2010).

Baseline PRL Ratings (Figure 3.1A). All baseline T2*-GRE phase image slabs were unwrapped, and T2*-GRE magnitude and phase images were averaged across echoes. All slabs (individual echo images and their average) were concatenated to generate whole-brain volumes. To register the T2*-GRE magnitude image to the qT1 map, the qT1 map was first cropped using FSL's *robustfov* tool to remove the neck region, which is not covered in the T2*-GRE acquisition. The qT1 map was then upsampled to 0.5 mm^3 resolution to better match the resolution of the T2*-GRE while minimizing resolution mismatch. The T2*-GRE magnitude scan was rigidly registered to the upsampled qT1

map using ANTs with sinc interpolation. The resulting transformation matrix was applied to preprocessed phase images.

All baseline WMLs were manually classified as PRL+ or PRL- using registered T2*-GRE magnitude, T2*-GRE unwrapped phase images, up-sampled T1 maps, and labelled baseline lesion masks. Lesions were excluded if PRL status could not be determined due to poor phase image quality, inhomogeneities, or abnormal veins. Only lesions fully visible within the concatenated supratentorial T2*-GRE field of view were included.

Quantitative WML Analysis: PRL+ vs PRL-. Native-resolution ($0.7mm$ isotropic) T1w images were segmented using PLAn 7T (Chapter 2.1), and individual WML segmentations were refined using the boundary correction pipeline (Chapter 2.2). Briefly, brain-extracted baseline T1w images were registered to each follow-up brain-extracted T1w image using affine registration. Then, baseline WML masks were split into individual lesion clusters, and each cluster was mapped to all follow-up images using the corresponding transformation matrix. For all timepoints (including baseline), each individual lesion was dilated into the surrounding normal appearing WM (NAWM), and lesion segmentations were corrected by fitting a bimodal Gaussian mixture model to the histogram of qT1 values within the dilated lesion mask. Lesion boundaries were refined by thresholding at the intersection of the two Gaussian curves. For each lesion, the median qT1 value and lesion volume were extracted for analysis. All boundary-corrected lesions were then combined and thresholded to produce a total lesion mask at all timepoints. Finally, the original GM, WM, and CSF segmentations at each timepoint were updated using the final, boundary-corrected WML masks.

Quantitative WML Analysis: Lesion Location and Lesion Layers. At baseline, each lesion was also classified as JC, PV, or DWM (Figure 3.1 C). To identify PV lesions, a ventricular segmentation mask was generated using FSL's *make_bianca_mask* tool, with the PLAn-derived CSF segmentation and the brain-extracted T1w image as input. A distance map from the ventricles was computed within the total WML mask using FSL's *distancemap* tool. This distance map was thresholded at $3mm$, binarised, and any lesion overlapping with this region was classified as PV. JC lesions were identified as lesions that overlapped with a dilated GM segmentation mask. Lesions that were neither JC nor PV were classified as DWM. Lesions that met both the JC and PV criteria

were excluded, as these were typically large confluent lesions and, as described in Chapter 2.2, more susceptible to noise and segmentation variability.

Lesions were also analysed by layer (Figure 3.1D). Layer masks were generated by iteratively eroding the lesion segmentation, with layers numbered from the outermost (layer 1) inward. Given that remyelination is most likely at the lesion edge, layers were grouped into 3 categories: layer 1, layer 2, and a combined *core* category for all deeper layers. Voxels in layer 1 adjacent to CSF or GM were excluded to minimize partial volume effects. Lesion layers were only included if all timepoints for a participant had corresponding layer data. Median qT1 values were extracted for each lesion layer. Since layer masks were defined by relative erosion, layer volumes are determined primarily by lesion size, shape, and partial voluming, and therefore were not considered a meaningful outcome for analysis.

3.1.2.3 Quantitative White Matter Lesion Analysis

Baseline values and longitudinal changes in qT1 and lesion volume were followed for PRL+ and PRL- (defined by the qualitative baseline ratings described in Section 3.1.2.2). Lesions smaller than $14mm^3$ or larger than $4000mm^3$ were excluded to reduce noise from small or likely confluent/diffuse WML. Analyses were repeated with datasets including lesion location and by lesion layer. For timepoints with clinical notes indicating the presence of enhancing lesions, the nearest 3T or 7T post-gadolinium scan was used to identify enhancing lesions. Lesions that showed gadolinium enhancement at any timepoint were excluded from the analysis.

3.1.2.4 Statistical Analysis

All median qT1 values were normalized by subtracting 1100 ms (the minimum observed value, rounded down) and then applying a log transformation. Similarly, all lesion volume measures were normalized by subtracting 13 mm^3 (just below the smallest lesion included) prior to log transformation.

Separate linear mixed-effects models (LMMs) were used to assess baseline and longitudinal differences in normalized lesion volume and qT1 values between PRL+ and PRL- lesions. Analyses were performed across: (i) all lesions to examine PRL main effects, (ii) lesion subgroups by anatomical location to assess PRL-by-location interactions, and (iii) lesion layers to assess PRL-by-layer

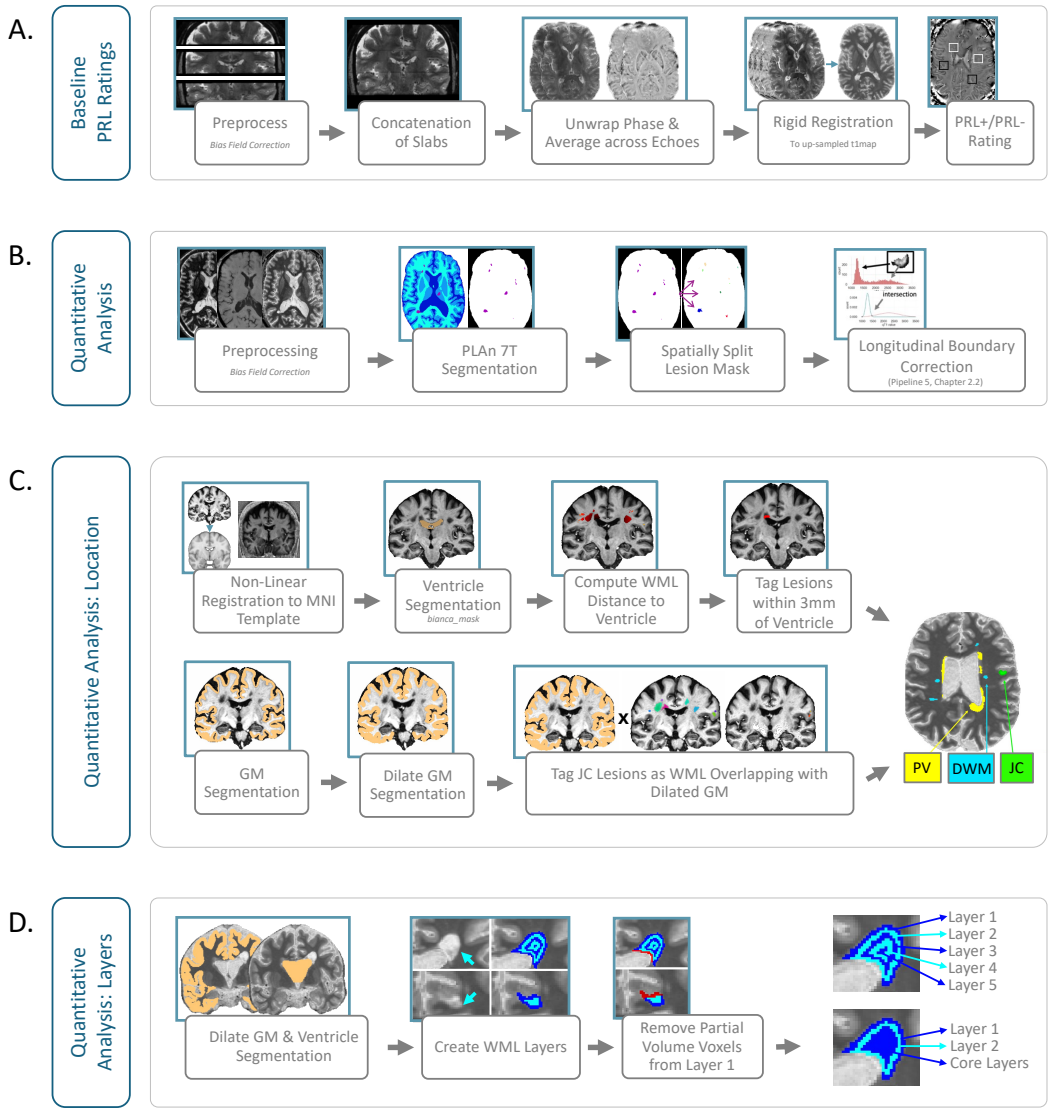


Figure 3.1: Overview of image processing steps for quantitative analysis of PRL+ vs PRL-. A) Baseline PRL ratings: T2*-GRE images underwent bias field correction, slab concatenation, phase unwrapping, echo-wise magnitude and phase averaging, and registration to the upsampled, preprocessed T1 map. B) To create longitudinal lesion masks, 7T MP2RAGE images were bias-field corrected and segmented into GM, WM, CSF, and WML using PLAn 7T. The WML segmentation was then spatially clustered to identify individual lesions and processed through the longitudinal boundary correction pipeline (see Chapter 2.2, Pipeline 5). C) To label WML as PV (top), a ventricular mask was generated via FSL's *make_bianca_mask*, using non-linear registration to MNI space and PLAn-derived CSF segmentation. Then, PV lesions were identified using a 3 mm thresholded distance map from the ventricular mask. Juxtacortical (JC) lesions were defined by overlap with a dilated PLAn GM segmentation. Remaining lesions were labeled as DWM. D) To generate lesion layers, GM and ventricle masks were dilated. One-voxel layers were iteratively created per lesion, excluding outer voxels overlapping with dilated GM or ventricle masks. Layers were classified from outermost (Layer 1) to innermost (core).

interactions. Each analysis was performed separately due to slight differences in the available datasets, as described in Section 3.1.2.2.

Random effects accounted for within-participant and within-lesion variability, as well as within-lesion layer correlation over time where appropriate. An

overview of all baseline and longitudinal models is provided in Table 3.1.

Table 3.1: Summary of linear mixed-effects models. Each row represents a distinct model. Fixed effects refer to the variables of interest used to explain variation in the dependent variable. Random effects account for repeated measures nested within participants and lesions, where specified. Covariates are shown in Appendix C.1.2

Model	**Norm. Dependent Variable**	**Dataset**	**Main Fixed Effect**	**Random Effects**
1	Lesion qT1	Baseline	PRL	1|ID
2	Lesion volume	Baseline	PRL	1|ID
3	Lesion qT1 slope	Longitudinal	PRL×age	1 + age|ID/Lesion
4	Lesion volume slope	Longitudinal	PRL×age	1 + age|ID/Lesion
5	qT1 (by location)	Baseline (Location)	PRL×location	1|ID
6	Volume (by location)	Baseline (Location)	PRL×location	1|ID
7	qT1 slope×location	Longitudinal (Location)	PRL×age×location	1 + age|ID/Lesion
8	Volume slope×location	Longitudinal (Location)	PRL×age×location	1 + age|ID/Lesion
9	qT1 by lesion layer	Baseline (Layers)	PRL×layer	1|ID/Lesion
10	qT1 slope×lesion layer	Longitudinal (Layers)	PRL×age×layer	1 + age|ID/Lesion/Layer

Abbreviations: norm = normalized; qT1 = quantitative T1; PRL = paramagnetic rim lesion; ID = participant ID

Model construction followed a stepwise approach. A full model was first fit, including the main fixed effect term and all candidate covariates. Covariates were considered based on theoretical relevance, prior literature, availability in the dataset, and to minimize collinearity with other predictors (Appendix C.1.2). Covariate selection was then performed by removing covariates with $p > 0.1$ to produce a reduced model. Deviations from normality and heteroscedasticity of model residuals were assessed using histograms, Q-Q plots, and residual-versus-fitted plots. Multicollinearity among fixed effects was evaluated using variance inflation factors calculated from a corresponding linear model with the same fixed effects. Model skewness and kurtosis were computed for both studentized and Pearson residuals to evaluate deviations from normality. If model residuals demonstrated substantial skew or kurtosis, outliers were identified as observations with standardized or Pearson residuals $\geq |3|$. Models were then refit after removing these observations as a sensitivity analysis.

An additional sensitivity analysis was performed for all LMMs after excluding voxels affected by partial voluming in WML segmentations (red voxels in Figure 3.1).

All models were fitted using the *R* package *lme4*. Fixed-effect and interaction estimates are reported as β coefficients with standard error (SE), t-values, and degrees of freedom (DoF) calculated using Satterthwaite's method. Model predictions, adjusted for covariates, were generated using the *ggeffects* package and are reported with their SE. For models including lesion location or lesion layer, post-hoc pairwise comparisons between PRL+ and PRL– lesions within each location or layer were performed. Group contrasts were computed using the *ggpredict* and *emmeans* functions from *ggeffects*, enabling location- or layer-specific comparisons. For pairwise contrasts, the estimated difference (Δ) is reported with its SE and p-value, and p-values were adjusted for multiple comparisons using the Tukey method.

3.1.3 Results

3.1.3.1 Participant Characteristics

Demographic and baseline MRI characteristics are summarized in Table 3.2. A total of 81 participants were included in the analysis, of whom 58 (71.6%) were female. At baseline, 60 participants (74.1%) were diagnosed with RRMS, 18 (22.2%) with SPMS, and 3 (3.7%) with primary progressive MS (PPMS). During the study period, 4 RRMS participants (4.9%) met the criteria for SPMS. Most participants were receiving DMTs during the MRI follow-up period, with 32 (39.5%) on high-efficacy treatments, 7 (8.6%) on moderate-efficacy treatments, 25 (30.9%) on low-efficacy treatments, and 17 (21.0%) untreated. The majority of participants self-identified as white and non-Hispanic (Table 3.2).

The average follow-up duration was 3.45 ± 1.6 years, with individual follow-up times up to 7.4 years between the first and last MRI scan. Participants underwent an average of 3.6 ± 1.9 MRI scans (range: 2 to 9). An overview of all longitudinal MRI scans included in the analysis is shown in Figure 3.2.

At baseline, the average number of PRL+ per participant was 5.4 ± 0.7 (Table 3.2). Overall, 19.8% had no detectable PRL+, 34.6% had 1-2 PRL+, and 37 participants, 45.7%, had 3 or more PRL+. Additionally, 5 participants had enhancing lesions (1-3 lesions) across baseline and/or follow-up scans, which were removed from the analysis. Across all participants, the plurality of WML

Table 3.2: Demographics and baseline characteristics of all participants. Continuous data are presented as mean ± standard deviation, and count data are presented with range [minimum, maximum].

		All Participants (n=81)
Participant Demographics	Age (years)	47.7±11.9
	Sex (Female, Male)	58, 23
	Sub-Diagnoses (RRMS, SPMS, PPMS)	60, 18, 3
	Treatment Efficacy (high, moderate, low, untreated, NA)	32, 4, 25, 17, 3
	Approximate age at onset (years)	32.8±10.3
	Race (White, Black/African, Asian, Other)	64, 14, 1, 2
	Ethnicity (Not Hispanic or Latino, Hispanic or Latino)	77, 4
Whole Brain MRI Characteristics	NAWM Volume (mm^3)	431619.4 ± 7814.4
	NAWM qT1 (ms)	1239.0 ± 39.6
	Total WML Volume (mm^3)	8485.3 ± 6765.9
	Total WML qT1 (ms)	1681.9 ± 153.4
Individual WML MRI Characteristics	# WML	59.2 ± 36.6 [3, 160]
	# PRL+	5.4 ± 6.1 [0, 30]
	# PRL-	54.1 ± 34.3 [1, 136]
	WML qT1 (ms)	1654.5 ± 305.1
	PRL+ qT1 (ms)	1834.7 ± 332.9
	PRL- qT1 (ms)	1637.5 ± 296.8
	WML Volume (mm^3)	119.2 ± 415.3
	PRL+ Volume (mm^3)	435.4 ± 1068.6
	PRL- Volume (mm^3)	89.3 ± 266.1

Abbreviations: NAWM = normal appearing white mater; qT1 = quantitative T1 times ; WML = white matter lesion ; PRL = paramagnetic rim lesion; RRMS = relapsing remitting MS; SPMS = secondary progressive MS; PPMS = primary progressive MS;

were PV for both PRL- and PRL+ lesions (36.7% and 34.3%, respectively). Among PRL- lesions, DWM lesions were more common (27.3%) than JC lesions (22.9%), whereas PRL+ lesions showed the opposite pattern, with more JC (24.2%) than DWM lesions (20.9%). The remaining 14% of all lesions (13.1% of all PRL- and 20.6% of all PRL+) were excluded from the analysis because they bordered both ventricles and cortex (likely confluent).

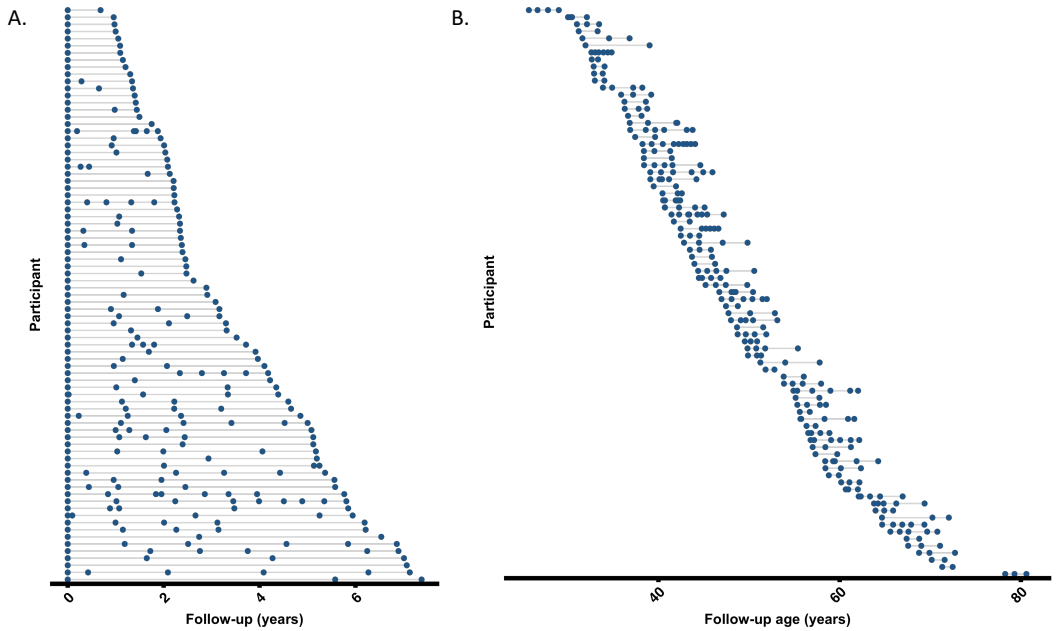


Figure 3.2: Longitudinal MRI scan timepoints for all study participants ($n = 81$). Each row represents a participant, and each point corresponds to the date of an MRI scan. A) All participants and their MRI scans, with time from baseline visit (in years) shown on the horizontal axis. B) All participant's age at the time of each MRI scan shown on the horizontal axis, ordered by baseline age.

3.1.3.2 Quantitative White Matter Lesion Analysis

Quantitative WML Analysis: PRL+ vs PRL-. In the baseline LMM, PRL+ lesions showed significantly higher normalized qT1 values (6.5 ± 0.02) compared to PRL– lesions (6.1 ± 0.02; $\beta = 0.3 \pm 0.03$, $t(4775) = 13.0$, $p < 0.001$), adjusting for baseline age. In this model, baseline age was not significantly associated with normalized qT1 ($\beta = 0.003 \pm 0.002$, $t(81) = 1.7$, $p = 0.09$). In a separate LMM model, PRL+ lesions had significantly greater normalized volume (4.8 ± 0.07) than PRL– lesions (3.5 ± 0.04; $\beta = 1.4 \pm 0.06$, $t(4775) = 21.0$, $p < 0.001$). Figure 3.3A and B show normalized lesion qT1 and volume for PRL+ and PRL– lesions. The findings for both models remained consistent in a sensitivity analysis excluding partial-volume voxels (Appendix C.1.2, Table C.1).

As depicted in Figure 3.3C, PRL+ and PRL– lesions showed significantly different normalized qT1 trajectories over time ($\beta_{\text{PRL}\times\text{age}} = 0.09 \pm 0.002$, $t(4325) = 3.6$, $p = 0.04$), after adjusting for normalized whole-brain qT1, which was also a significant covariate ($\beta = 0.6 \pm 0.04$, $t(6451) = 15.0$, $p = 0.04$). PRL+ lesions exhibited a significant increase in normalized qT1 over time (slope $= 0.007 \pm 0.003$, $p = 0.05$), whereas PRL– lesions did not significantly change over time (slope $= -0.001 \pm 0.003$, $p = 0.6$).

In the LMM of normalized lesion volume over time (Figure 3.3D), lesion volume increased significantly for both PRL+ and PRL– lesions (PRL–: slope $= 0.02 \pm 0.002$, $p < 0.001$; PRL+: slope $= 0.02 \pm 0.005$, $p < 0.001$), but the difference in slopes between PRL+ and PRL– lesions was not statistically significant ($\beta_{\text{PRL}\times\text{age}} = 0.006 \pm 0.005$, $t(4115) = 1.2$, $p = 0.2$). Across all timepoints, PRL+ lesions were larger than PRL– lesions ($\beta_{\text{PRL}} = 1.4 \pm 0.06$, $t(4628) = 22.0$, $p < 0.001$). This model was adjusted for participant's total follow-up time ($\beta = 0.03 \pm 0.02$, $t(61) = 1.5$, $p = 0.1$), approximate age at disease onset ($\beta = -0.03 \pm 0.04$, $t(61) = -2.5$, $p = 0.01$), and normalized whole-brain qT1 ($\beta = 0.3 \pm 0.09$, $t(990) = 2.9$, $p = 0.004$).

Figures 3.3E and F show spaghetti plots of normalized qT1 and lesion volume over age for individual lesions. For visualization, the number of PRL– lesions was randomly downsampled to match the number of PRL+ lesions. An example PRL+ lesion with increasing qT1 over time is highlighted in Figure 3.3G.

For both longitudinal models, removal of statistical outliers (to reduce residual kurtosis) and excluding partial-volume voxels did not change the findings: all significant fixed effects remained consistent in both direction and statistical significance (Appendix C.1.2, Appendix Tables C.2 and C.3).

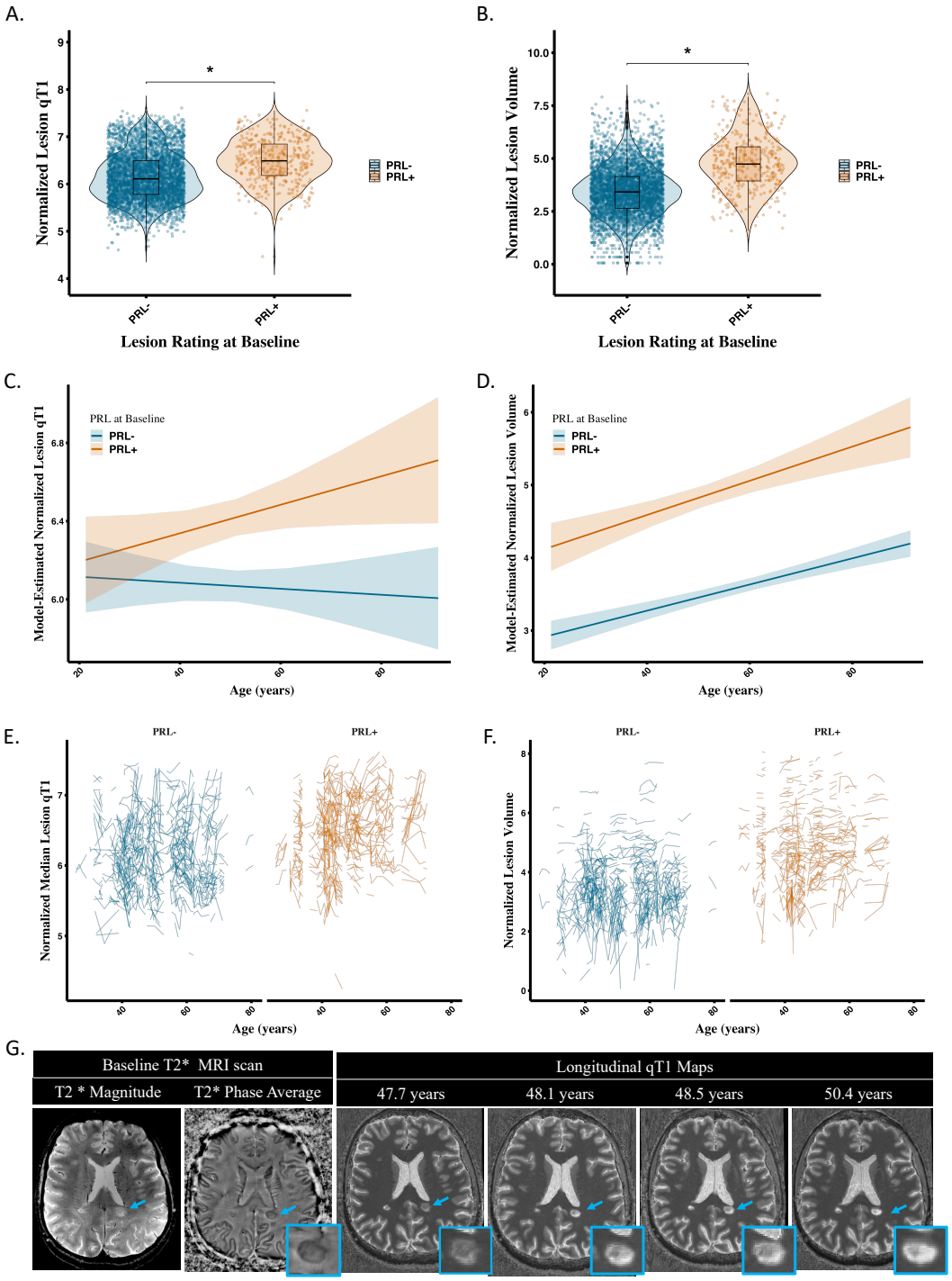


Figure 3.3: Baseline normalized lesion (A) qT1 and (B) volume for PRL+ and PRL– lesions. Estimated longitudinal changes from mixed-effects models: (C) normalized qT1 median and (D) volume by age at MRI. Individual lesion trajectories shown as spaghetti plots of (E) normalized qT1 medians and (F) volumes; PRL– lesions were randomly downsampled to match the number of PRL+. (G) Example PRL+ lesion (left: baseline T2*-GRE magnitude and unwrapped phase; right: longitudinal qT1 maps) showing an increase in qT1 over time.

Quantitative WML Analysis: Lesion Location. At baseline, PRL+ lesions had higher normalized qT1 than PRL- lesions across all locations (Figure 3.4A).

This difference was largest in PV lesions ($\Delta = 0.35 \pm 0.04$, $p < 0.001$), followed by DWM ($\Delta = 0.27 \pm 0.05$, $p < 0.001$) and JC lesions ($\Delta = 0.20 \pm 0.05$, $p < 0.001$). Among PRL- lesions, qT1 was significantly lower in DWM compared to JC ($\Delta = -0.36 \pm 0.02$, $p < 0.001$) and PV lesions ($\Delta = -0.40 \pm 0.02$, $p < 0.001$), but the difference between JC and PV lesions did not reach significance ($\Delta = -0.04 \pm 0.02$, $p = 0.07$). Among PRL+ lesions, all pairwise differences between locations were significant: qT1 in DWM was lower than in JC ($\Delta = -0.29 \pm 0.07$, $p < 0.001$) and PV lesions ($\Delta = -0.48 \pm 0.06$, $p < 0.001$), and JC lesions had lower qT1 than PV lesions ($\Delta = -0.19 \pm 0.06$, $p = 0.004$). These findings were consistent even after partial volume voxels were excluded (Appendix C.1.2, Table C.4).

At baseline, PRL+ lesions were larger than PRL– lesions across all lesion locations (Figure 3.4B), with the largest difference observed in PV lesions ($\Delta = 1.24 \pm 0.09$, $p < 0.001$), followed by DWM ($\Delta = 1.14 \pm 0.12$, $p < 0.001$) and juxtacortical lesions ($\Delta = 0.92 \pm 0.12$, $p < 0.001$). These PRL+ vs PRL– differences remained significant after removing voxels affected by partial voluming (Appendix C.1.2 Table C.4). While some within-PRL group differences between lesion locations were observed (Figure 3.4B), they were not statistically significant after partial voluming voxels were excluded (Appendix C.1.2 Table C.4).

In the longitudinal LMM for normalized qT1 by PRL status and lesion location, there was a significant overall difference in age-related qT1 change between PRL+ and PRL- lesions ($\beta_{\text{PRL}\times\text{age}} = 0.01 \pm 0.005$, $t(3455) = 2.5$, $p = 0.01$), with PRL+ lesions showing a steeper increase.

Among PRL- lesions, DWM lesions showed a significant decrease in qT1 over time ($\Delta = -0.012 \pm 0.003$, $p < 0.001$), whereas JC ($\Delta = 0.004 \pm 0.003$, $p = 0.23$) and PV lesions ($\Delta = -0.005 \pm 0.003$, $p = 0.10$) did not significantly change over time. Within lesion location group, PRL+ qT1 slopes were not significantly different than 0 (DWM: $\Delta = 0.001 \pm 0.006$, $p = 0.91$; PV: $\Delta = 0.002 \pm 0.005$, $p = 0.66$), though JC lesions showed a trend toward increasing qT1 over time ($\Delta = 0.010 \pm 0.005$, $p = 0.061$).

Direct comparison of slopes between PRL+ and PRL– lesions within each location revealed a significant difference in DWM lesions ($\Delta_{\text{PRL- vs. PRL+}} = -0.013 \pm 0.005$, $p = 0.012$) and in PV lesions ($\Delta_{\text{PRL- vs. PRL+}} = -0.007 \pm 0.004$, $p = 0.05$). This reflects that qT1 decreased over time in PRL– lesions but was slightly increasing in PRL+ lesions. Differences between PRL+ and PRL-

slopes were not statistically significant in JC lesions ($\Delta = -0.006 \pm 0.005$, $p = 0.18$). Overall, these findings suggest that longitudinal qT1 trajectory differences between PRL+ and PRL– lesions are primarily driven by DWM and PV lesions.

In the longitudinal location model of lesion volume, all lesions showed an increase in normalized volume over time (Figure 3.4). However, the rate of volume change did not significantly differ between PRL+ and PRL- lesions ($\beta_{\text{PRL}\times\text{age}} = 0.002 \pm 0.01$, $t(3273) = 0.2$, $p = 0.8$). Post hoc pairwise comparisons confirmed no significant PRL-related differences within any location ($p > 0.1$). In contrast, lesion location significantly influenced longitudinal volume trajectories regardless of PRL status. This effect was primarily driven by PRL- lesions: normalized volume increased more steeply in DWM compared to JC lesions ($\Delta = 0.014 \pm 0.003$, $p < 0.001$). No significant slope differences by location were found among PRL+ lesions (all $p > 0.98$).

For both longitudinal models, removal of statistical outliers (to reduce residual kurtosis) and excluding partial-volume voxels did not change the findings: all significant fixed effects remained consistent in both direction and statistical significance (Appendix C.1.2, Appendix Tables C.5 and C.6).

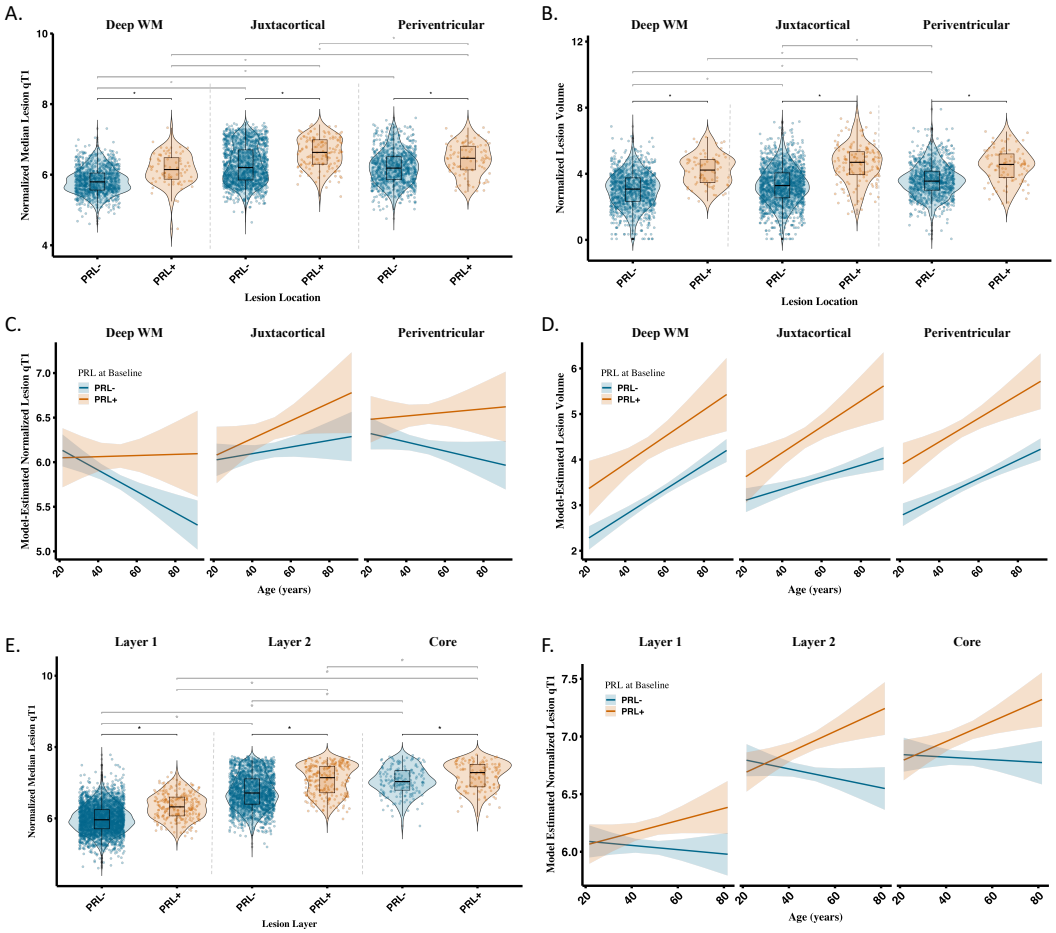


Figure 3.4: PRL+ vs. PRL- lesions by lesion location and layer at baseline and longitudinally. A) Baseline normalized lesion qT1 and B) volume for PRL+ and PRL- lesions in deep white matter, juxtacortical, and periventricular locations. C) Predicted normalized qT1 slopes from linear mixed-effects models showing longitudinal changes in PRL+ vs. PRL- lesions by location. D) Predicted normalized lesion volume slopes from linear mixed-effects models showing longitudinal changes in PRL+ vs. PRL- lesions by location. E) Baseline normalized qT1 in lesion layers: outermost (layer 1), second outermost (layer 2), and core. F) Predicted normalized qT1 slopes from linear mixed-effects models showing longitudinal changes in PRL+ vs. PRL- by lesion layer.
Black lines with * indicate significant post-hoc pairwise differences between PRL+ and PRL- lesions, adjusted for multiple comparisons.
Grey lines with * indicate significant post-hoc pairwise differences within PRL groups.

To assess qT1 within lesion layers, lesions were divided into three layers: layer 1 corresponding to the outer edge, layer 2 to the next inward layer, and all remaining voxels grouped as the *core*. At baseline, pairwise post-hoc tests showed that PRL+ lesions had higher normalized qT1 compared to PRL– lesions across all layers ($p < 0.001$, Figure 3.4E). Within both PRL– and PRL+ lesions, inner layers exhibited higher qT1 than outer layers (see Appendix C.1.1 for full post-hoc summary).

This baseline LMM was adjusted for several covariates: normalized lesion volume, the presence of EL, and normalized whole-brain median qT1. Both

normalized lesion volume ($\beta = 0.09 \pm 0.005$, $t(3829) = 18.2$, $p < 0.001$) and normalized whole-brain qT1 median ($\beta = 0.8 \pm 0.2$, $t(85) = 4.6$, $p < 0.001$) were positively associated with lesion-layer qT1. Presence of enhancing lesions showed a trend toward lower lesion-layer qT1 ($\beta = -0.2 \pm 0.1$, $t(99.8) = -1.9$, $p = 0.06$).

A linear mixed model was fitted to assess longitudinal changes in normalized lesion qT1 across layers by PRL status. As described previously (Table 3.1), the random-effects structure initially included `(age | ID/Lesion/Layer)`. Near-zero variance at the layer level led to a singular fit, so the structure was simplified to `(age | ID/Lesion)`, retaining random intercepts and slopes for lesions nested within participants, while removing layer-level random effects.

Median layer qT1 in PRL+ lesions showed a significantly different trajectory than layers in PRL- lesions ($\beta_{\text{PRL}\times\text{age}} = 0.007$, $p < 0.001$, Figure 3.4F). Across all layers, PRL+ increased over time compared to PRL- lesions (post hoc PRL- vs. PRL+ Δ, $p < 0.001$).

Only layer 2 and the Core layer in PRL+ significantly increased over time (PRL+ layer 2 Slope:$0.009 \pm 0.004, p = 0.004$; PRL+ core layer Slope:$0.009 \pm 0.004, p = 0.007$). In PRL-, pairwise post hoc tests showed that layer 2 decreased more steeply over time compared to layer 1 ($\Delta = -0.002 \pm 0.004, p < 0.001$) and compared to Core ($\Delta = -0.003 \pm 0.004, p = 0.001$). For PRL+, Layer 2 and the core layer increased more steeply than Layer 1 ($\Delta = 0.004 \pm 0.001, p = 0.001$ and $\Delta = 0.003 \pm 0.001, p = 0.02$, respectively).

The LMM controlled for normalized median whole-brain qT1, normalized lesion volume, treatment group, and scanner. Whole-brain median qT1 and baseline lesion volume were strongly associated with lesion-layer qT1 ($p < 0.001$), while treatment and scanner effects were not significant.

3.1.4 Discussion

In this 7T longitudinal natural history study, the evolution of PRL+ and PRL- was investigated in a large cohort of 81 participants with up to seven years of follow-up. PRL+ lesions were larger and exhibited higher qT1 values at baseline and showed steeper increases in qT1 over time, supporting their characterization as more destructive both cross-sectionally and longitudinally. These findings suggest that therapies aimed at controlling inflammation at the rims of PRL+ might impact lesion evolution. While previous studies have linked PRLs to disease progression and tissue damage in smaller cohorts or over shorter

follow-up periods, the present findings extend this evidence to a larger population and longer timescale. Moreover, novel insights into the spatial dynamics of lesion evolution are provided. DWM and PV PRL+ lesions demonstrated progressive tissue damage compared to PRL- lesions. Among PRL- lesions, only DWM lesions demonstrated decreasing qT1 longitudinally, which may indicate possible increase capacity for partial tissue recovery for PRL- in these regions. Finally, longitudinal analysis across lesion layers revealed that qT1 increases in PRL+ lesions occur both at the lesion edge (layer 2) and core, highlighting pathological progression throughout the lesion.

Building on prior work, our findings demonstrate that PRL+ lesions continue to evolve over long-term follow-up, showing sustained increases in qT1 compared to PRL- lesions. This aligns with the notion that chronic MS lesions may undergo dynamic pathological processes rather than follow a uniform trajectory (Calvi et al., 2022). The use of 7T qT1 mapping, which is sensitive to both demyelination and, at qT1 higher values, axonal loss, provides a sensitive marker of progressive tissue damage within these lesions (Kolb et al., 2021). Our results are consistent with prior studies that reported greater tissue disruption in PRL+ lesions over shorter intervals using diffusion and magnetization transfer imaging. For instance, Elliott et al. (2023) observed that PRL+ lesions showed reductions in nMTR, nT1, and FA, alongside increased RD over 1.3 years, changes that are associated with decreases in myelin and/or axonal density or integrity (Granziera et al., 2021). Similarly, Weber et al. (2022) reported lesion enlargement and increased ADC, reflecting expansion of the extracellular space due to demyelination and axonal loss, in PRL+ lesions up to three years post-gadolinium enhancement. Our study extends these findings, showing that PRL+ lesions not only accumulate damage over time but continue to expand across longer follow-up periods, supporting previous observations of sustained lesion growth in PRL+ (Absinta et al., 2019; Blindenbacher et al., 2020).

Interestingly, in contrast to previous studies reporting that PRL– lesions tend to decrease in volume over time (Absinta et al., 2019), this cohort showed a significant longitudinal increase in PRL– volume. This discrepancy may be partly explained by differences in study design. Absinta et al. (2019) analysed a small, matched subset of 27 PRL– lesions specifically selected for comparison with PRL+ lesions, whereas the present analysis included thousands of

PRL– lesions across a large cohort, capturing a broader and more heterogeneous spectrum of lesion behaviour. While it is also important to consider methodological influences, such as the potential for systematic bias in the image processing pipeline, the test–retest analysis (Chapter 2.2) did not reveal any tendency toward artefactual volume increases at follow-up. Indeed, other studies have highlighted heterogeneity in PRL- lesion evolution, reporting that while an overall reduction in lesion size may be observed, a subset of PRL- lesions still expand over time (Dal-Bianco et al., 2021; Klotz et al., 2025; Weber et al., 2022). Follow-up analyses on this dataset could help characterize this variability in PRL- size evolution by applying categorical definitions of lesion growth similar to those used by Dal-Bianco et al. (2021) and Absinta et al. (2019) to differentiate PRL- lesions that expand from those that remain stable or shrink. This would enable further investigation into whether specific lesion characteristics, such as anatomical location or baseline volume, are associated with PRL- trajectories and could reveal important subtypes of PRL-.

While PRL- lesions showed a significant increase in volume over time in our results, their mean qT1 values remained stable. The lesion layer analysis revealed stable qT1 values across all PRL- layers, including the lesion edge, where early repair processes would typically be expected. This stability suggests that the observed decrease in mean qT1 likely reflects peripheral lesion expansion into voxels with lower qT1, rather than tissue repair. Future analyses could test this hypothesis by directly examining the relationship between volume and qT1 trajectories, for example, assessing whether lesions with the largest volume increases also show the greatest decreases in qT1. Additionally, combining categorical analyses of both volume and qT1 changes may help classify lesions into distinct patterns of evolution, such as those that are improving (decreasing in both qT1 and volume) versus those that are progressively destructive (increasing in both). This multidimensional approach could provide a more nuanced understanding of PRL- evolution.

PRL+ lesions were most commonly classified as PV, followed by JC and then in DWM, consistent with a prior 3T study showing higher PRL+ frequency in PV and JC regions (Guo et al., 2021). However, a larger 3T study with 254 MS patients reported that PRL+ were more commonly found in DWM than in JC regions (Meaton et al., 2022), which contrasts with our 7T findings. This discrepancy may stem from improved lesion conspicuity at the cortex-WM boundary at 7T, making JC PRLs easier to detect than at lower field strengths.

Alternatively, the identification of JC lesions using a dilated gray matter ribbon may have over-classified lesions near, but not touching, the cortex, potentially inflating the frequency of JC PRLs.

PV and JC lesions also showed higher qT1 and larger volumes compared to DWM, supporting prior findings linking PV WML to greater tissue injury and clinical severity. For example, annual relapse rate was positively associated with PV PRL+ burden (Guo et al., 2021). PRL- in DWM exhibited the steepest qT1 decrease over time, which may reflect greater repair potential. This aligns with neuropathological evidence that remyelination is most impaired near the ventricles and improves with distance (Tonietto et al., 2023), suggesting DWM lesions, which are further away from ventricular CSF, may be more permissive to endogenous repair. In contrast, JC regions are more exposed to meningeal inflammatory infiltrates, which could hinder repair in PRL- (Magliozzi et al., 2007).

The lesion-layer analysis provides further insight into lesion pathology. As expected, the innermost lesion layer (core) had the highest qT1 and the outer layer (layer 1) the lowest, reflecting increasing tissue damage toward the centre of the lesion. PRL+ lesions consistently showed significantly higher qT1 than PRL- in every layer, with the largest PRL- vs PRL+ difference in the inner edge layer (layer 2). This pattern suggests accentuated peripheral damage in PRL+. Longitudinal qT1 trajectories were similar in inner and outer layers, indicating that tissue damage progresses both at the lesion edge and in the core over time. This supports the notion that chronic active lesions are not limited to rim-confined pathology but involve ongoing degeneration throughout the lesion. Histopathological studies align with this interpretation, demonstrating that iron-positive microglia and macrophages accumulate at the lesion rim, where they drive ongoing demyelination, while the lesion core shows evidence of axonal transection and dystrophy. It is worth noting that lesion layers were generated independently at each timepoint, which may introduce variability if lesion segmentations differ across visits. Applying a consistent lesion mask across timepoints might improve sensitivity to detect subtle layer-specific changes Elkady et al. (2025). Nonetheless, the qT1 layer findings remain consistent with a smouldering, centrally propagating process in PRL+ lesions, where both edge and core undergo progressive degeneration beyond the initial inflammatory rim.

There are several limitations to this study that present opportunities for future investigation and additional analyses. First, we fit linear models to lesion trajectories, which may oversimplify the true underlying dynamics. Prior studies have suggested that lesion evolution can be non-linear, particularly in chronic active lesions such as PRLs (Weber et al., 2022). Second, lesion age was not explicitly accounted for in our analyses. Reductions in lesion volume and shortening of qT1 relaxation times have been reported within the first year following lesion formation (Absinta et al., 2016), suggesting that some younger lesions may still be in a post-acute recovery phase. Analysing lesions by their estimated age (e.g., time since gadolinium enhancement on MRI) could help clarify whether observed changes differ between early and chronic phases. As only a few ELs were detected in this dataset, there was an insufficient number of EL to support their analysis as a distinct group. However, a retrospective review of earlier 1.5T and 3T scans within our natural history cohort could enable identification of lesion onset, facilitating future analyses by lesion age.

Additionally, the evolution of the paramagnetic rim itself was not tracked, and PRL+ lesions were identified solely on the basis of their baseline appearance. However, not all paramagnetic rims persist, and emerging evidence suggests that rim resolution may have prognostic value. For example, resolving PRLs are associated with slower disability progression and lower rates of PIRA, independent of relapse activity (Reeves, Bartnik, Jakimovski, Mohebbi, Bergsland, Salman, Schweser, Wilding, Weinstock-Guttman, Dwyer and Zivadinov, 2024). Future work will leverage the rim visualization tool developed in Chapter 2.3 to longitudinally assess rim status and identify PRLs that resolve over time. This will allow analysis of qT1 and volume trajectories before and after rim disappearance and comparison of PRLs with persistent rims versus those that resolve (Elkady et al., 2025; Weber et al., 2022). Finally, while the current study did not include disability outcomes, future work will link lesion-level imaging markers to longitudinal clinical data (e.g., EDSS, cognitive decline) to evaluate their relevance for disease monitoring.

In summary, this study provides a comprehensive in-vivo characterization of chronic MS lesion evolution over extended follow-up. PRL+ lesions demonstrated more severe and progressive tissue damage than PRL- lesions, both cross-sectionally and longitudinally, and across spatial dimensions within the lesion. These findings are consistent with a chronic inflammatory pathophysiology, as previously described (Absinta et al., 2016a) and highlight the value

of 7T qT1 mapping of PRL+ and PRL- lesions as a sensitive tool for tracking tissue damage and as a potential biomarker of more aggressive disease.

3.2 Effect of Anakinra on Chronically Inflamed White Matter Lesions in Multiple Sclerosis

In the previous section (Section 3.1), I presented findings from a 7T natural history study, demonstrating that chronic PRL+ lesions follow destructive trajectories in both qT1 and volume. Building on this, the current section focuses on 7T MRI analyses from a clinical trial evaluating the therapeutic effects of anakinra, an interleukin-1 receptor antagonist, on chronic MS lesions. I first provide background on the role of interleukin-1β in MS pathophysiology and the rationale for anakinra as a potential intervention. I then describe the trial design and analysis for 7T MRI WML endpoints. The analyses examine whether anakinra affects the paramagnetic rim status of chronic WMLs or modifies the qT1 and volume trajectories of PRL+ lesions relative to PRL– lesions, which are interpreted in the context of the natural history trajectories presented in the preceding section.

3.2.1 Overview of Interleukin-1β and Anakinra in Multiple Sclerosis

Interleukin-1 beta (IL-1β), a member of the interleukin-1 (IL-1) family, is a potent pro-inflammatory cytokine that mediates various immune responses and is implicated in chronic inflammatory conditions. In MS and experimental autoimmune encephalomyelitis (EAE), an established animal model of MS that reproduces many immunopathological features of central nervous system (CNS) inflammation and demyelination in MS, IL-1β has been shown to be expressed primarily by activated microglia, but also in macrophages in response to inflammatory stimuli (Burm et al., 2016; Herx et al., 2000; Prins et al., 2013). The binding of IL-1β to its receptors, specifically IL-1 receptor type I (IL-1R1) and IL-1 receptor type III (IL-1R3), triggers a cascade of inflammatory responses that activate both innate and adaptive immune mechanisms. In the adaptive immune response, IL-1β promotes recruitment and activation of T-cells and B-cells. IL-1β activates innate immune responses by stimulating microglia and

astrocytes, which release inflammatory cytokines and recruit polymorphonuclear cells (PMNs) into the CNS, leading to blood-brain barrier breakdown and demyelination (Prins et al., 2013). Indeed, IL-1β has direct neurotoxic effects: IL-1β, especially in combination with TNF-α, induces oligodendrocyte death and impairs astrocyte glutamate uptake, promoting excitotoxic injury. IL-1β also drives the conversion of astrocytes to a neurotoxic *A1* phenotype (in concert with TNF-α and complement C1q) (Kamma et al., 2022). These actions contribute to chronic demyelination and axonal injury in MS (Kamma et al., 2022).

Several studies have associated IL-1β with disease progression in MS patients. Levels of IL-1β in the CSF are elevated in MS compared to controls (Dujmovic et al., 2009) and are correlated with higher cortical lesion volume and number, but not with the presence of acute gadolinium-enhancing lesions (Seppi et al., 2014). In an early study, increased IL-1β expression was found in chronic active lesions compared to acute lesions (Cannella and Raine, 1995). CSF levels of IL-1β in MS participants have also been shown to be related to higher disability. In a study of 170 RRMS patients, detection of IL-1β in the CSF was associated with higher progression index and Multiple Sclerosis Severity Scale scores (MSSS) after four years of follow-up, compared to patients with undetectable levels of IL-1β, who had a higher likelihood of having a benign MS phenotype (Rossi et al., 2014). Additionally, FDA-approved DMTs for MS have been shown to decrease IL-1β in CSF (DMT: natalizumab) (Mellergard et al., 2010) and in monocytes and serum (DMT: interferon-beta and glatiramer acetate) (Burger et al., 2009; Guarda et al., 2011).

In animal models, IL-1β is required for full disease: mice lacking IL-1R1 or IL-1β are substantially protected from EAE (Levesque et al., 2016; Matsuki et al., 2006). Preclinical animal studies using EAE have also provided evidence that IL-1β antagonists can provide clinical improvement. One study demonstrated that while injections of IL-1α can exacerbate disease severity, the administration of soluble IL-1 receptor as an IL-1 antagonist leads to a significant reduction in clinical paralysis and wasting compared to controls (Jacobs et al., 1991). Other studies have shown that in-vivo treatment with recombinant human IL-1 receptor antagonist suppressed clinical symptoms associated with EAE (Martin and Near, 1995) and reduced T-cell proliferation (Badovinac et al., 1998). These findings suggest that interrupting IL-1β can dampen the pathological immune response in MS.

Despite the growing body of evidence supporting the involvement of IL-1β in MS pathogenesis, the inhibition of this cytokine has not been clinically investigated in the context of MS treatment. Anakinra, a recombinant, non-glycosylated form of the human IL-1 receptor antagonist, has been developed to block the biological activity of IL-1β by competitively inhibiting its binding to IL-1R1. While it was approved in 2001 for rheumatoid arthritis and autoinflammatory syndromes and has a well-characterized safety profile (Lopalco et al., 2016), anakinra has never been trialed in MS.

Given the evidence from preclinical studies demonstrating clinical improvement with the use of IL-1β antagonists, and evidence that ventricular anakinra delays the repopulation of microglia (Bruttger et al., 2015), there is a compelling argument for exploring the therapeutic effects of anakinra in MS. By targeting IL-1β, anakinra may inhibit the proliferation of activated macrophages and microglia at the edges of chronic active lesions, thereby mitigating the inflammatory processes that contribute to neurodegeneration.

In this section, I aim to explore the effects of anakinra on PRLs as visualized with 7T MRI from the clinical trial *Anakinra for the Treatment of Chronically Inflamed White Matter Lesions in Multiple Sclerosis*. As part of the ATaC-MS trial's radiological outcomes, I describe the assessment of whether up to 12 weeks of subcutaneous anakinra alters the presence or characteristics of the paramagnetic rim, influences the evolution of qT1 in chronic MS lesions, or shows evidence of clinical treatment response.

3.2.2 Methods

3.2.2.1 Overview of Clinical Trial Design

This study includes data from the ATaC-MS clinical trial (Appendix A.1), a single-site, open-label, single-arm Phase 2a study conducted at the NIH clinical centre. The trial enrolled eight participants, five of whom completed the full 24-week protocol. Participants were recruited through self-referral, the NIH Clinical Center Recruitment Office, and the NIH Clinical Research Volunteer Program, or through ongoing Neuroimmunology Clinic protocols at the NIH. Eligible individuals had progressive or stable MS, radiological evidence of chronic active lesions in the WM, and no new lesion formation in the preceding 3 months or clinical relapse in the preceding 12 months, as determined

from off-protocol MRI scans. The trial's full inclusion and exclusion criteria are provided in Appendix C.3.

A schematic overview of the study design and data collection is shown in Figure 3.5. After providing informed consent, participants were trained to self-administer subcutaneous injections of anakinra. Participants began anakinra treatment at week 0 with a daily dose of 100 mg. After the initial dose, participants received a follow-up phone call within 24-28 hours to assess concerns or monitor adverse reactions. The treatment period lasted for 12 weeks with two dose escalations: at week 4, the dose was escalated to 200mg/day, and at week 8, to the target dose of 300mg/day. Week 12, four weeks on the target dose, marked the final on-treatment study visit. The drug was then discontinued, and participants were monitored during a 12-week off-treatment safety monitoring period, including follow-up visits 4 and 12 weeks after drug discontinuation (week 16 and week 24). All study visits were in-person and included a 7T MRI (see Appendix A.3.2 for MRI acquisition parameters) and clinical assessment.

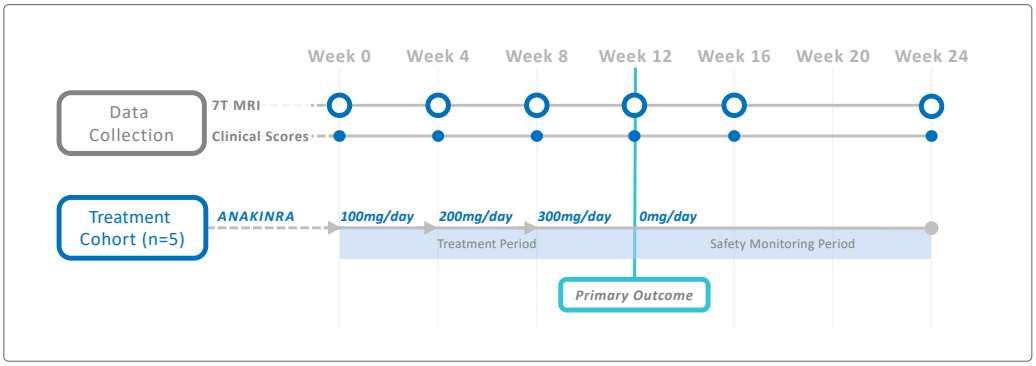


Figure 3.5: Overview of study design. 7T MRI scans and clinical disability scores were collected at baseline and every 4 weeks. Anakinra dosing was escalated every 4 weeks during the treatment period, reaching a target dose of 300 mg/day by Week 8. The primary outcome was assessed at Week 12 (highlighted in cyan blue).

3.2.2.2 Clinical Trial Endpoints

The primary outcome of the ATaC-MS trial was the proportion of PRL+ lesions per participant that exhibited rim disappearance following the 12-week dosing period. The null hypothesis was defined as no rim disappearance and would be rejected if at least one lesion rim disappeared.

The study was powered based on prior data showing that paramagnetic rims rarely change over short periods, with an estimated change rate of only 1% (0.01) in untreated patients (Absinta et al., 2021). Using a binomial exact test,

the probability of observing at least one participant with a lesion change out of 5 total was calculated to be less than 5% ($p < 0.05$). This meant that if one or more participants showed a change, the result would be considered statistically significant. The study had 80% power to detect a true lesion change rate of 27% or higher across participants, making even a single observed paramagnetic rim disappearance meaningful.

Secondary and exploratory endpoints included: (1) qualitative evaluation of the persistence or change in rim appearance across all available timepoints; (2) longitudinal changes in qT1 relaxation times within PRL+ lesions compared to PRL- lesions; (3) changes in lesion volume for PRL+ lesions relative to PRL- lesions throughout the trial; and (4) changes in clinical scores.

3.2.2.3 MRI Image Processing

Figure 3.6 provides an overview of T2*-GRE and T1w image processing steps for radiological WML ratings and quantitative analysis. All T2*-GRE and T1w contrasts underwent bias-field correction except the estimated T1 map, which was directly reconstructed on the scanner from the MP2RAGE images with bias-field correction.

7T MRI data were processed for baseline radiological ratings (Figure 3.6, dark blue) and longitudinal radiological ratings (Figure 3.6, light blue). First, all T2*-GRE phase image slabs were unwrapped. Magnitude and phase images from both echo times were averaged across echoes, and all slabs (individual echo images and their average), were concatenated to generate whole-brain volumes. These concatenated T2*-GRE images were then intensity-normalized to a 0-1 scale using a 98th percentile normalization:

$$I_{\text{norm}} = \frac{I - \min(I)}{P_{98}(I) - \min(I)} \tag{3.1}$$

where $P_{98}(I)$ is the 98th percentile of the image intensity matrix I. This can be mathematically defined as

$$P_{98}(I) = \inf\{x \in \mathrm{R} \colon F_i(x) \geq 0.98\} \tag{3.2}$$

where $F_i(x)$ is the empirical cumulative distribution function (CDF) of the image intensities I.

For within-session registration of T2*-GRE images to the corresponding T1 map, T1 maps were first cropped to remove the neck region using FSL's *robustfov* tool and then upsampled to 0.5 mm isotropic resolution. T2*-GRE volumes were subsequently rigidly registered to the T1 map with sinc interpolation using ANTs' registration tool (Avants et al., 2008).

For longitudinal ratings, all follow-up scans were registered to the upsampled baseline T1 map. First, longitudinal T1 maps were rigidly registered to baseline, and then the registration matrices from within-session T2*-to-T1 map registration and longitudinal T1 map follow-up-to-baseline registration were concatenated and applied to longitudinal T2*-GRE and phase images using sinc interpolation. Finally, skull-stripped, upsampled T1 maps were normalized to a 0–1 scale using mean-based normalization. This followed the same formula as Equation (3.1), but used the image mean (mean(I)) instead of the 98th percentile ($P_{98}(I)$). Mean-based normalization was chosen for the T1 maps because, unlike T2*-GRE images, they were skull-stripped and thus less affected by non-brain intensity outliers that could distort normalization based on high-percentile values.

For quantitative analysis (Figure 3.6, purple), T1w images (in their native $0.7mm$ isotropic resolution) were segmented using PLAn 7T (Chapter 2.1). Individual WML segmentations were created using the boundary correction pipeline (Chapter 2.2). An additional step was added to the pipeline: after the baseline segmentation was extracted from PLAn 7T and spatially split into individual lesions using FSL's *cluster* tool, lesion splitting was manually corrected as needed and reviewed by a neurologist (M.I.G.). The resulting baseline WML mask was used as the starting point for the longitudinal boundary correction pipeline (Chapter 2.2). Finally, the original NAWM segmentations at each timepoint were corrected using the final, boundary-corrected WML masks.

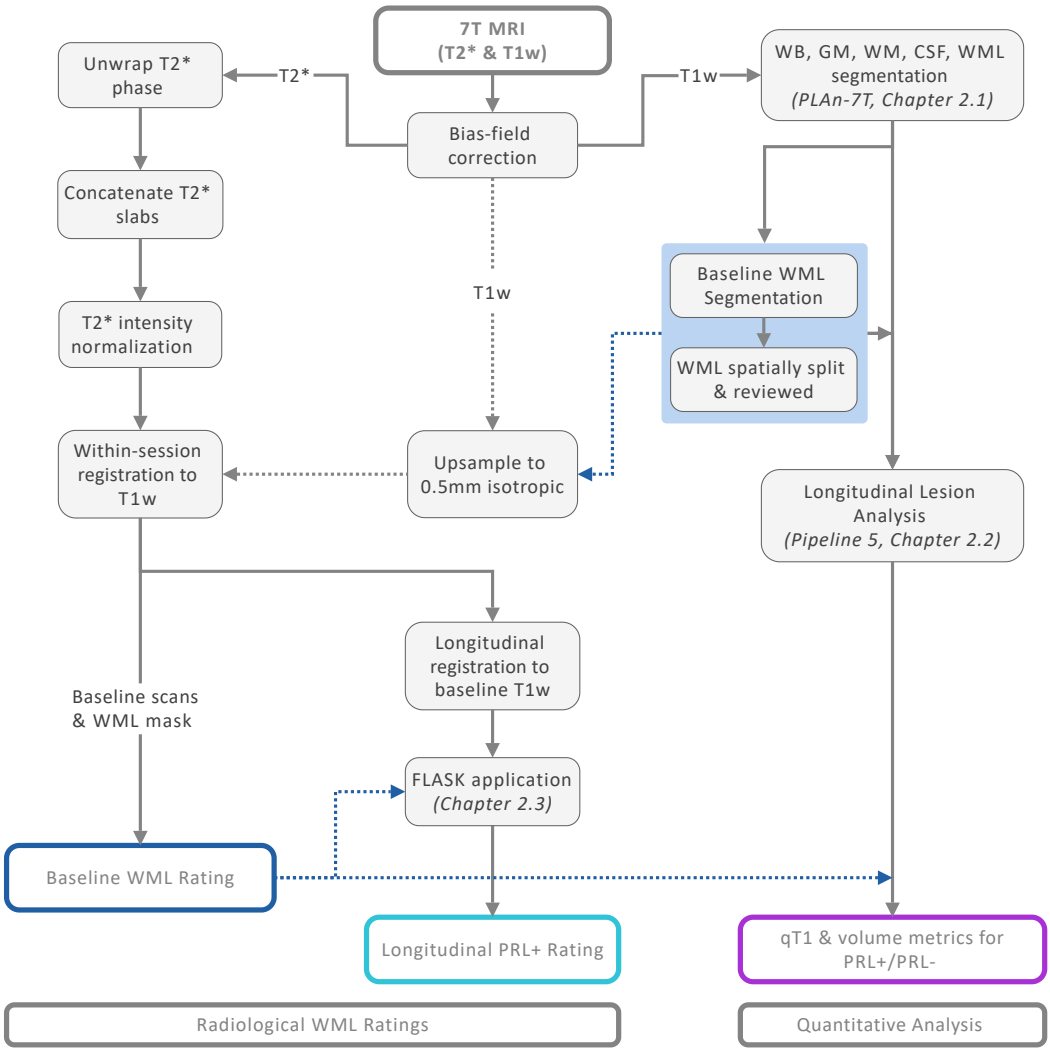


Figure 3.6: Overview of image processing steps. 7T MRI (T1w and T2*-GRE) processing steps for radiological WML Ratings (Baseline and Longitudinal Ratings) and Quantitative Analysis. Longitudinal PRL+ Ratings (cyan blue) are image processing steps required for the primary outcomes.

3.2.2.4 Radiological Assessment of the Paramagnetic Rim

Baseline Ratings of Paramagnetic Rim Status

Two neurologists (D.S.R. and M.I.G.) separately rated all WML at baseline as having a paramagnetic rim (PRL+) or not having a paramagnetic rim (PRL-). Hereafter, the expert raters are referred to as Rater A and Rater B, with anonymized labels assigned arbitrarily and with no fixed correspondence to D.S.R. or M.I.G. The specific assignment was retained internally to enable calculation of intra-rater reliability, which is described in subsequent sections. For the baseline PRL rating, within-session registered T2* magni-

tude images, T2*-GRE unwrapped phase images, up-sampled T1 maps and the baseline WML mask with individual lesions labelled were provided. Lesions were excluded from the analysis under specific conditions, such as when the PRL status could not be determined due to the phase image's quality, the presence of inhomogeneities, or the presence of abnormal veins. Additionally, only lesions entirely visible on the concatenated supratentorial T2*-GRE field of view were included. Any discrepancies in ratings between the two raters were discussed and resolved.

Longitudinal Ratings of Paramagnetic Rims status. All lesions identified as PRL+ at baseline were longitudinally evaluated to determine whether the rim signal evolved over time. To address different study questions, two distinct sets of longitudinal ratings were conducted: Primary Outcome Ratings and Secondary Outcome Ratings.

Primary Outcome Ratings. This analysis was designed to address the primary clinical trial endpoint: the effect of anakinra on PRL resolution. For this reason, the rating only included participants who completed the trial and compared PRL appearance between week 0 and week 12, the timepoint when participants had escalated to the full dose. The goal here was to provide a blinded and reliable assessment of whether treatment was associated with PRL change.

To achieve this, a custom Flask-based application (Chapter 2.3) was adapted to facilitate blinded longitudinal ratings (Figure 3.7A). For each PRL+ lesion, axial slices were presented across two anonymized columns corresponding to different timepoints. Each lesion was shown twice: once with the columns arranged in ascending chronological order (week 0 to week 12) and once in reverse chronological order (week 12 to week 0), to blind raters to the true timepoint ordering. All PRL+ lesion pairs from completers were pooled and randomly distributed across web pages, with each page displaying eight lesions in randomized order. Two independent raters then assessed changes in rim appearance, blinded to both participant identity and timepoint order. This design minimized rater bias and ensured an objective evaluation of treatment effects. These design choices supported the reliability and validity of the primary endpoint measurement. Lesions were displayed with multiple MRI contrasts (estimated T1 maps, average T2*-GRE magnitude, T2*-GRE phase at echo 2, and averaged T2*-GRE phase, all registered to baseline structure). Rating

options included: *rim more conspicuous*, *rim less conspicuous*, *no change*, *rim disappeared*, *rim appeared*, and *other*, with space for free-text comments.

Lesions with inter- or intra-rater discrepancies were compiled into a dedicated webpage for consensus review using the same Flask-based interface and presentation format. Finally, all lesions rated as showing change were further reviewed in a second consensus session using the full, registered NIfTI images to confirm the presence and nature of rim evolution.

*Secondary Outcome Ratings.*The second analysis was designed to address a broader, exploratory question: whether any change in the paramagnetic rim occurred across any available timepoint. For this reason, it included all participants, both completers and those who withdrew, ensuring that all available data were examined for potential rim changes. Because the dataset was unevenly structured (e.g., missing timepoints in participants who withdrew; Example 3.7B, bottom) and the goal was exploratory, this analysis did not use the same blinding steps as the primary outcome rating. A separate Flask-based application (Chapter 2.3) was developed to evaluate PRL changes across up to six available timepoints, displayed in chronological order in a single row (Figure 3.7B). Each lesion appeared once, with lesions scrambled across participants to reduce systematic bias. The rating form included assessment fields for rim changes across three intervals: week 0–12, week 12–24, and week 0–24. As with the primary rating, lesions with inter- or intra-rater discrepancies were compiled into a consensus Flask application for joint review, and any lesions marked as changing were further verified using the full registered NIfTI images.

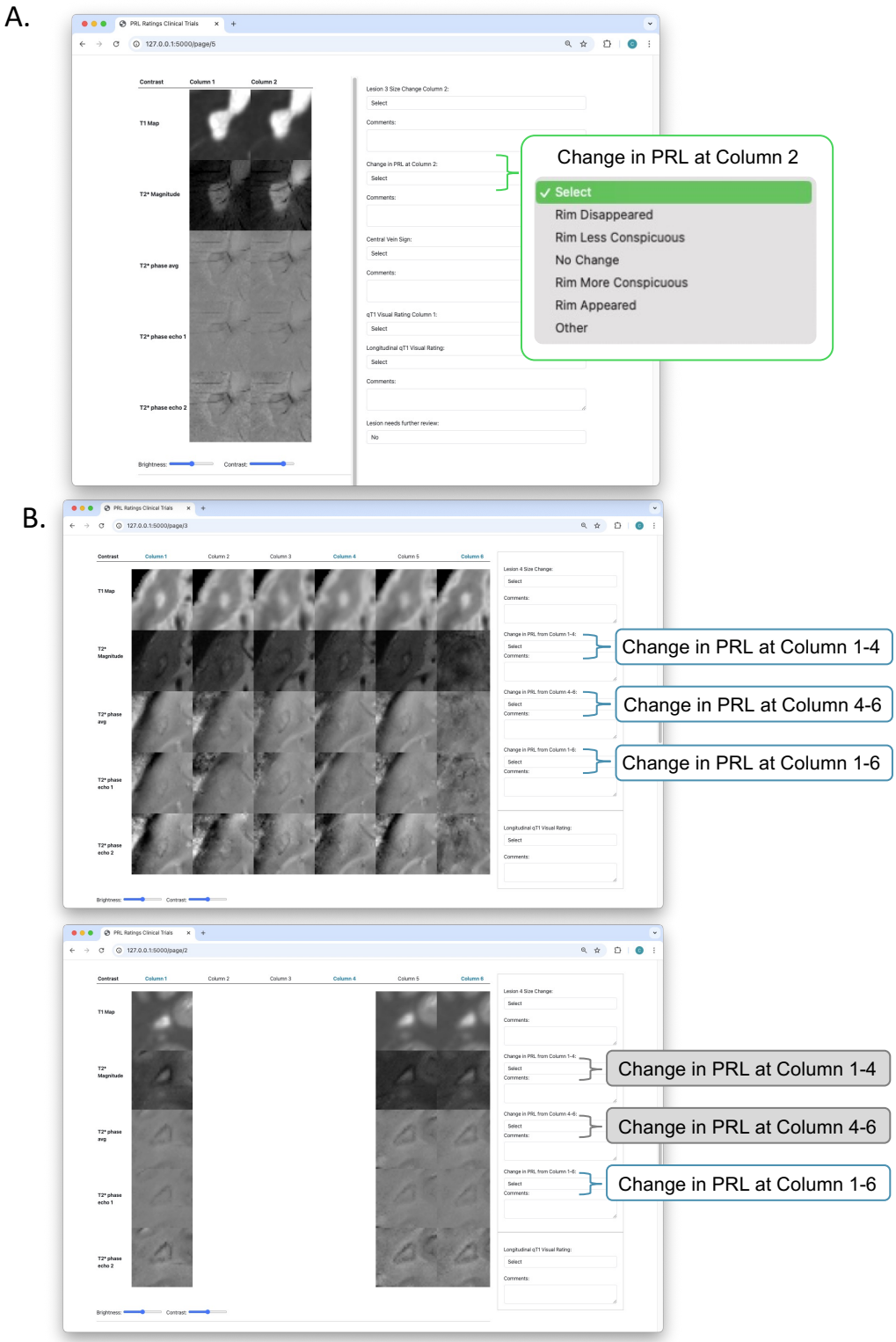


Figure 3.7: Example Flask applications for Primary and Secondary PRL Ratings. A) Primary outcome rating: longitudinal change is assessed between two timepoints (Week 0 and Week 12). Columns are anonymized, and the order may be chronological or reverse chronological to enable blinded assessment. B) Secondary outcome rating: two example cases. Top, a participant who completed the trial with all six timepoints; longitudinal changes are rated for Week 0–12, Week 12–24, and Week 0–24 intervals (highlighted in blue). Bottom/right, a participant who withdrew early and is missing intermediate timepoints; only the Week 0–24 interval can be rated. This rating includes all available participants and timepoints to ensure that any PRL changes are captured, even when data are incomplete.

3.2.2.5 Quantitative White Matter Lesion Analysis

In the native resolution of the T1 maps (0.7mm isotropic), the median qT1 time and volume were extracted for each individual lesion, NAWM, and the total lesion mask across all available timepoints. Timepoints exhibiting motion artifacts in the T1 maps, as determined by visual inspection, were excluded from the analysis. Longitudinal changes in qT1 and lesion volume were examined for individual lesions classified at baseline as either PRL+ or PRL- within the T2*-GRE field of view. In addition, lesions identified as PRL+ and rated as showing rim change in the longitudinal assessments were qualitatively compared to other lesions within the same participant to explore whether rim evolution was associated with distinct qT1 or volume dynamics.

3.2.2.6 Statistical Analysis

Inter- and Intra-Reliability Calculations of PRL Ratings. Inter- and intra-reliability calculations of PRL ratings. Inter- and intra-rater reliability for baseline and longitudinal PRL ratings, pooling lesion ratings across all participants were assessed. At baseline, inter-rater reliability between two raters was calculated using an unweighted Cohen's κ. The raters assessed the presence of PRL (yes/no) for all WML in the baseline T2*-GRE field of view.

For the primary outcome ratings, both raters assessed the change of rim status of each PRL+ twice, once in chronological order and once in reverse chronological order. Inter-rater reliability was calculated for the two sets of ratings: rating 1 (chronological order) and rating 2 (reverse chronological order). Intra-rater reliability for chronological and reverse chronological ratings was calculated after unscrambling reverse chronological order ratings. Each rating was assigned a value from 0 to 5, corresponding to PRL+ change categories: *rim more conspicuous*, *rim less conspicuous*, *no change*, *rim disappeared*, and *rim appeared*. For the secondary outcome ratings, inter-rater reliability was calculated for each available rating. Cohen's κ values were interpreted according to established guidelines: ≤ 0 indicating *No Agreement*, 0.01 - 0.20 indicating *None to slight*, 0.21 - 0.40 indicating *Fair*, 0.41 - 0.60 indicating *Moderate*, 0.61 - 0.80 indicating *Substantial*, and 0.81 - 1.00 indicating *Almost Perfect* agreement.

Longitudinal Changes in Rim Status of PRL+. The total number of lesions exhibiting rim change across the entire cohort and the proportion of

rim-changing lesions per participant were reported.

Quantitative Analysis of Lesion qT1 and Volume. All median qT1 values and volume measures were normalized as described previously (Section 3.1.2.4).

At baseline, differences in normalized lesion qT1 and volume between PRL+ and PRL- lesions were assessed using separate LMMs with lesions nested within participants. PRL status was included as a fixed effect. To assess longitudinal changes in lesion qT1 and volume between PRL+ and PRL- lesions, I collaborated with a statistician (T.W.) to appropriately model the data. Although LMM were initially considered, they did not adequately fit the data and failed to meet key model assumptions. As a result, a repeated-measures ANCOVA (rm-ANCOVA) was used, treating time as a categorical variable. Separate models were fit for normalized qT1 and normalized lesion volume, each including fixed effects for PRL status at baseline, timepoint, and their interaction, while adjusting for selected covariates. Participant-wise covariates considered in the design of the statistical models included sex, baseline age, baseline normalized NAWM volume, baseline normalized NAWM median qT1, baseline normalized total lesion burden, and baseline normalized total lesion qT1 median. Lesion-wise covariates considered included baseline normalized lesion volume and baseline normalized qT1 median. Fixed effects were retained if they were not collinear and significantly contributed to the model, with significance defined as $p < 0.1$. Within-group time effects for PRL+ and PRL- lesions were assessed using sliced effects. Pairwise comparisons between timepoints within each PRL group were conducted and adjusted for multiple testing using the Tukey-Kramer method. Analyses were repeated with and without statistical outliers, defined as observations with studentized residuals exceeding ± 3.

Missing data resulting from motion artifacts were imputed using a group-level approach. For each PRL group and time interval, the percent change from the previous timepoint was calculated and used to adjust the lesion's prior value to estimate the missing data point.

Changes in Clinical Scores. Clinical test scores were obtained at baseline and follow-up timepoints, provided participants were willing. For timepoints with missing scores, the missing value was imputed as the average of the participant's scores from the previous and future timepoints. If the last timepoint was missing a score, the previous timepoint's result was carried forward. After testing for normality, changes in disability scores over time were analysed

using one-way ANOVA. Pairwise comparisons were performed between weeks using the Friedman test or the Holm-Sidak test for multiple comparisons with repeated measures.

3.2.3 Results

3.2.3.1 Participant Characteristics

A total of eight participants enrolled in the clinical trial (mean age = 49.2 ± 11 years; six female). While all participants had a diagnosis of RRMS at the time of enrolment, they were clinically stable, with no new lesions in the three months prior to enrolment and no clinical relapses in the preceding twelve months. The mean baseline EDSS score was 2.0 ± 1.5, with scores ranging from 0 to 3.5 across all participants. Five participants completed the full dosing period (completers), while three withdrew from the trial (non-completers); one after the baseline visit and two after the week 4 visit. All three participants who withdrew from the study subsequently completed both required safety monitoring visits.

Table 3.3 summarizes demographic and baseline quantitative imaging and baseline clinical measures for both completers and non-completers. At baseline, the number of WMLs varied by participant, ranging from 20 to 77 among trial completers. All participants had more PRL- compared to PRL+ at baseline. Across all participants' WML, 23% were radiologically determined as PRL+.

Table 3.3: Participant demographics, whole brain MRI characteristics, lesion-level MRI characteristics, and clinical scores for completers and non-completers. Data shown as mean ± SD, with ranges [min, max] where applicable.

		Completers(n=5)	**Non-Completers (n=3)**
Participant Demographics	Age (years)	50±11	48±2.2
	Sex (Female, Male)	3, 2	3, 0
	Sub-Diagnoses (RRMS, SPMS, PPMS)	5, 0, 0	3, 0, 0
	Race (White, Black/African)	5, 0	2, 1
	Ethnicity (Not H/Lat, H/Lat)	4, 1	3, 0
Whole Brain MRI Characteristics	NAWM Volume (mm^3)	385420±69544	385060±77274
	NAWM qT1 (ms)	1197±17.1	1208±50.7
	Total WML Volume (mm^3)	8800±8516	10108±13145
	Total WML qT1 (ms)	1688±106	1693±297
Individual Lesion MRI Characteristics	# WML	22.5±25.7 [20, 77]	28.2±28.5 [17, 83]
	# PRL+	5.2±1.6 [3, 8]	8.7±6.4 [1, 13]
	# PRL–	40±26.7 [17, 73]	48±28.6 [16, 71]
	WML qT1 (ms)	1602±246	1638±305
	PRL+ qT1 (ms)	1864±386	1746±313
	PRL- qT1 (ms)	1570±202	1619±225
	WML Volume (mm^3)	112±174	110±244
	PRL+ Volume (mm^3)	261±297	79.5±76.9
	PRL- Volume (mm^3)	93.2±143	116±262
Clinical Scores	EDSS	2.0±1.5 [0, 3.5]	1.7±1.5 [0, 3.0]
	Scripps	90±9.8 [75, 100]	92±8.5 [83, 100]
	SDMT	50±3.7 [45, 54]	53±22.7 [36, 79]
	PASAT*	48±21.3 [10, 59]	37±33.2 [13, 60]
	9HPT DH	19±2.7 [15, 22]	20±2.9 [18, 23]
	25 FTW	4.9±0.8 [4.2, 6.1]	6.4±7.7 [4.9, 7.7]
	TUG	6.9±1.4 [5.0, 8.6]	9.2±3.4 [6.7, 13]

Abbreviations: NAWM = normal appearing white matter; qT1 = quantitative T1 times; WML = white matter lesion; PRL = paramagnetic rim lesion; RRMS = relapsing-remitting multiple sclerosis; SPMS = secondary progressive multiple sclerosis; PPMS = primary progressive multiple sclerosis; H/Lat = Hispanic / Latinex; EDSS = Expanded Disability Status Scale; Scripps = Scripps Neurological Rating Scale; SDMT = Symbol Digit Modalities Test; PASAT = Paced Auditory Serial Addition Test; 9HPT DH = 9-Hole Peg Test, dominant hand; 25 FTW = 25-Foot Timed Walk; TUG = Timed Up and Go. *One completer did not complete PASAT at baseline.

3.2.3.2 Radiological Assessment of the Paramagnetic Rim Analysis

Across all enrolled participants ($n = 8$), a total of 402 WML within the baseline T2*-GRE field of view were independently evaluated for the presence of a paramagnetic rim (PRL+ vs. PRL-) by the two blinded raters. Inter-rater agreement was substantial, with an unweighted Cohen's κ of 0.63 (Table 3.4). Following consensus review, 58 lesions were confirmed as PRL+ at baseline, 32 of which were from the five trial completers.

All clinical trial visits from completers ($n = 5$) were included in the primary outcome rating assessing the change in paramagnetic rim. Inter-rater reliability for rim status change was poor, with Cohen's weighted kappa indicating none to slight agreement: $\kappa = -0.19$ for chronologically ordered lesions and $\kappa = 0.057$ for reverse chronological presentations (Table 3.4). For both ratings, five lesions were rated differently between raters, out of 27 lesions for rating 1 and 25 lesions for rating 2. Although the number of disagreements was small, the majority of lesions were rated *no change*, creating a highly skewed distribution. The weighted Kappa accounts for agreement expected by chance, which is very high when most ratings fall in a single category, so even a few disagreements can substantially reduce the Kappa value (Bexkens et al., 2018). This may explain why inter-rater reliability appears low despite only a small number of differing ratings. Intra-rater reliability between rating 1 and rating 2 was *fair* for Rater A ($\kappa = 0.31$) and *moderate* for Rater B ($\kappa = 0.47$, Table 3.4).

The secondary outcome ratings assessed whether any participant exhibited a change in PRL status at any timepoint. For this rating, lesions were presented once in chronological order. For trial completers, raters evaluated changes across three intervals: week 0-12, week 12-24, and week 0-24. For non-completers, raters assessed change over all available timepoints from week 0-24 (Figure 3.8). Inter-rater reliability remained poor for the week 0-12 ratings ($\kappa = -0.06$) but showed slight improvement for week 12-24 ($\kappa = 0.29$) and week 0-24 ($\kappa = 0.12$) ratings (Table 3.4).

Table 3.4: Inter- and Intra-Rater Reliability of Radiological Paramagnetic Rim Lesion Ratings.

Description	**Type of Reliability**	**Sample size***	**Cohen's Kappa details**	**Cohen's Kappa**
Baseline Ratings (Completers & Withdrawn)				
Classification of all lesion segmentations at baseline as PRL+ or PRL-.	Inter-rater	396	Unweighted, z=12.7	0.633
Longitudinal: Primary Outcome Rating (Completers only)				
Change in paramagnetic rim from Week 0 – Week 12 in PRL+ lesions only.	Rating 1: Inter-rater	27	Quadratic weighted, z=-1.2	-0.19
	Rating 2: Inter-rater	25	Quadratic weighted, z=0.38	0.057
	Rater A: Intra-rater	25	Quadratic weighted, z=-2.65	0.47
	Rater B: Intra-rater	28	Quadratic weighted, z=-1.68	0.31
Longitudinal: Secondary Outcome Ratings (Completers & Withdrawn)				
Change in paramagnetic rim from Week 0 – Week 12 in PRL+ lesions only. (Completers only)	Inter-rater	26	Quadratic weighted, z=-0.37	-0.061
Change in paramagnetic rim from Week 12 – Week 24 in PRL+ lesions only. (Completers only)	Inter-rater	25	Quadratic weighted, z=1.77	0.29
Change in paramagnetic rim from Week 0 – Week 24 in PRL+ lesions only. (Completers & Non-Completers)	Inter-rater	40	Quadratic weighted	0.12

Abbreviations: PRL+ = paramagnetic rim lesions
* Lesions marked as *other* in any rating were excluded from reliability analyses. These lesions were reviewed during consensus ratings.

As described in the methods Section 3.2.2.4, all lesions with intra- or inter-rater discrepancies were reviewed together by both raters in a dedicated consensus Flask application. Any lesion showing a change in rim status was then confirmed using the original NIfTI images. The resulting final longitudinal PRL rating provided a single, agreed-upon score for each lesion, which was used to assess changes in paramagnetic rim status over time (Table 3.5).

During consensus reviews, six lesions initially labelled as PRL+ at baseline were reclassified as *not rateable* and removed from further analysis. These lesions appeared to have a paramagnetic rim in the dedicated MRI viewer but

were inconspicuous in the Flask app used for trial ratings, preventing reliable evaluation. An overview of PRL+ lesions removed from the analysis is provided in Appendix C.1.

The summary of rim status changes across longitudinal ratings is presented in Table 3.5. A change in PRL status between week 0 and week 24 could not be assessed in 11 lesions due to motion-related artifacts or because the lesion was not visible in the T2*-GRE field of view at all follow-up timepoints. In the first longitudinal rating assessing the primary outcome, 1 of the 25 PRL+ lesions evaluated demonstrated a change in rim appearance, with the rim rated as less conspicuous at week 12 compared to baseline (Figure 3.8A). The per-participant proportion of lesions with rim changes was 0.13 for one participant and 0 for all other participants. Since no rims fully disappeared, the primary outcome was not met, and the trial was considered negative. The second longitudinal rating, which included all time points, confirmed this single rim change between week 0 and week 12. No additional rim changes were observed at any other timepoints across all completers and non-completers.

Table 3.5: Overview of Rim Changes in PRL+ from Week 0–Week 12 in Trial Completers and Week 0–24 in All Participants

Outcome	**Week 0–12 (Primary Outcome)**					**Week 0–24 (Secondary Outcome)**							
Total # PRL+ Lesions	25					40							
Total # of *Rim Disappeared*	0					0							
Total # of *Rim Less Conspicuous*	1					1							
Participant	A002	A003	A004	A005	A008	A002	A003	A004	A005	A008	A001	A006	A007
Total # PRL+ Lesions	8	5	3	4	6	8	3*	3	4	6	4*	1	12
Total # of *Rim Disappeared*	0	0	0	0	0	0	0	0	0	0	0	0	0
Total # of *Rim Less Conspicuous*	0	1	0	0	0	0	1	0	0	0	0	0	0
Per participant proportion of rims disappeared	7	5	3	4	6	7	3	3	4	6	4	1	12
Total # of *Rim Less Conspicuous*	1/8	0/5	0/3	0/4	0/6	1/8	0/3	0/3	0/4	0/6	0/4	0/1	0/12

Abbreviations: PRL+ = paramagnetic rim lesions
Excludes WMLs that could not be rated due to motion artifacts or because they were not visible in the follow-up T2-GRE field of view.

The lesion whose paramagnetic rim became less conspicuous between week 0 and week 12 was further examined. Compared to other PRL+ lesions within the same participant, this lesion had a similar normalized qT1 value (Figure 3.8B) but a higher-than-average normalized lesion volume (Figure 3.8C). Over time, there was no clear trend in normalized qT1 or volume. However, this lesion showed a modest positive slope of normalized volume over time, which aligned with the general trajectory observed across all PRL+ lesions in the participant. The lesion also exhibited a slight decrease in qT1 between Weeks 16 and 24; however, the magnitude of change falls within the range expected from test-retest variability (Chapter 2.2) and may not reflect true biological change.hello

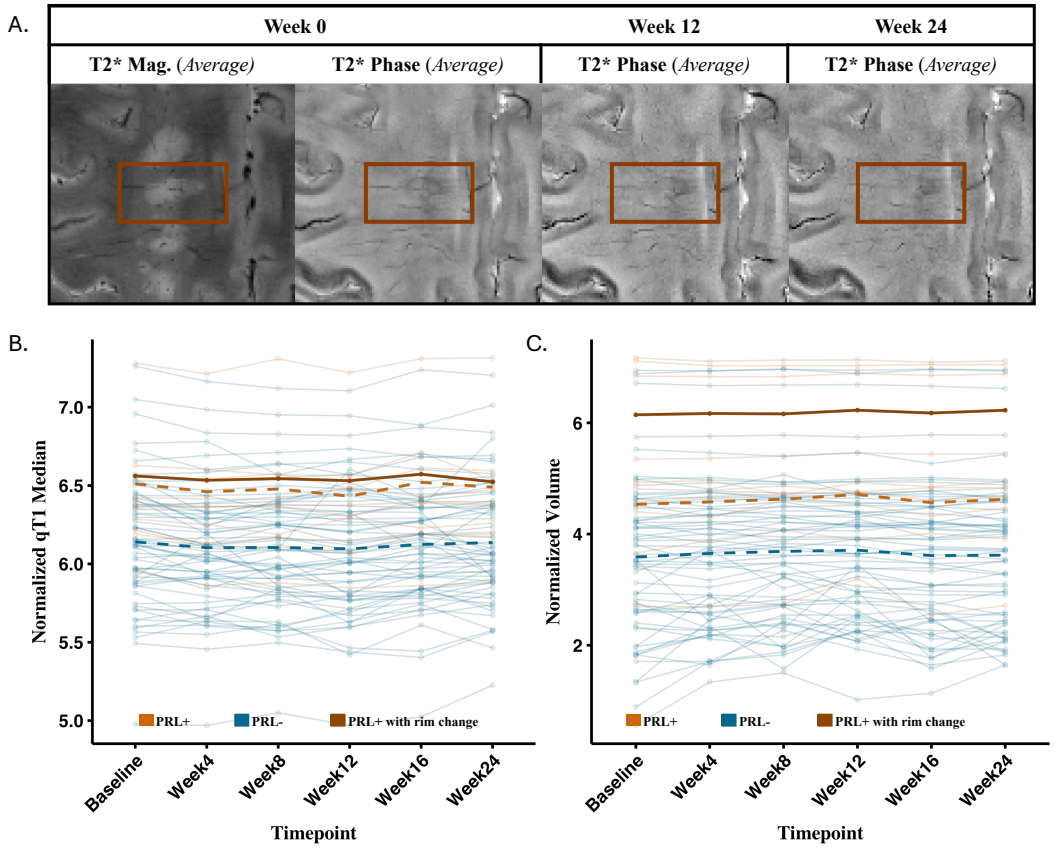


Figure 3.8: PRL+ with reduced rim conspicuity between Week 0 – Week 12. A) Example axial slices showing the lesion on preprocessed baseline T2*-GRE magnitude and the lesion's change in rim appearance on baseline, registered week 12 and week 24 phase images. B) Normalized qT1 and C) Normalized Volume over time for the lesion with rim change (solid dark orange line), compared to all PRL+ (light orange) and PRL- (blue) lesions from the same participant.

3.2.3.3 Quantitative White Matter Lesion Analysis

Normalized lesion volume and normalized qT1 values were examined in PRL+ and PRL- lesions, as well as their evolution over time among trial completers ($n = 5$). For this analysis, only lesions larger than 14 mm^3 were included, resulting in 24 PRL+.

Baseline nested LMM controlling for baseline normalized total lesion qT1 median and normalized baseline lesion volume, showed that PRL+ lesions had significantly higher normalized qT1 ($\beta = 1.71$) compared to PRL- lesions ($\beta = 1.49$, $p < 0.001$, Figure 3.9A). Covariates were positively associated with normalized lesion median qT1 (baseline total lesion qT1: $\beta = 0.54$, $p = 0.013$; baseline lesion volume: $\beta = 0.15$, $p < 0.001$). Additionally, at baseline, PRL+ lesions had greater volume ($\beta = 0.76$) compared to PRL- ($\beta = -3.36$, $p = 0.0019$, Figure 3.9B) while controlling for baseline normalized lesion qT1 median.

Longitudinal changes of normalized lesion qT1 and lesion volume were assessed among PRL+ and PRL- lesions using RM-ANCOVA. Two participants' T1 maps at week 8 and week 12 were excluded from the analysis due to motion artefacts, and data were imputed as previously described in the methods (section 3.2.2.6).

The qT1 model estimates, adjusted for normalized baseline NAWM qT1 and total lesion volume, are shown in Figure 3.9C. A significant main effect of PRL status was observed ($F(1, 228) = 22.44$, $p < 0.0001$), with PRL+ lesions consistently exhibiting higher qT1 values compared to PRL- lesions across all timepoints. Neither the main effect of time ($F(5, 887) = 1.51$, $p = 0.19$) nor the interaction between time and PRL status ($F(5, 887) = 0.49$, $p = 0.7835$) reached significance. However, simple effects tests confirmed that PRL+ lesions had significantly higher qT1 values than PRL- lesions at each individual timepoint (all $p < 0.0001$), suggesting stable differences of PRL+ vs PRL- lesions over the 24-week period.

Among PRL- lesions, unadjusted pairwise comparisons suggested modest decreases in qT1 from baseline to later timepoints (week 4, 8, 12, 16); however, none of these comparisons remained statistically significant after Tukey-Kramer correction (Appendix C.8). To examine the impact of extreme values, a sensitivity analysis was conducted by excluding nine outlier observations (three single-timepoint lesions and one lesion measured at all six timepoints). After exclusion, normalized baseline qT1 decreases from baseline to weeks 4 and 12 reached statistical significance (Appendix C.9).

PRL+ lesions were significantly larger than PRL- lesions across all timepoints (main effect of PRL status: $F(1, 227) = 40.66$, $p < 0.0001$, Figure 3.9D). There was no significant main effect of time ($F(5, 1145) = 1.27$, $p = 0.2735$) and no significant interaction between PRL status and time ($F(5, 1145) = 0.70$, $p = 0.62$). Tests of simple effects confirmed that PRL+ lesions remained significantly larger than PRL- lesions at each timepoint (all $p < 0.0001$). Within the PRL- group, a small but significant volume decrease was observed between baseline and week 4 ($t(1145) = 3.75$, $p_{\text{adj}} = 0.01$). No significant volume changes were detected over time within the PRL+ group ($F(5, 1145) = 0.33$, $p = 0.90$), indicating stable lesion volume. All post-hoc comparisons are reported in Appendix C.10; no statistical outliers were identified in this model.

To assess the impact of imputed data, RM-ANCOVA models were re-run using only the three participants (1091 individual lesions) with complete time-

points. Results were consistent with the full dataset, showing stable differences in lesion volume and qT1 between PRL+ and PRL- lesions over time, a slight decrease in qT1 among PRL- lesions from baseline to follow-up timepoints, and a small volume reduction between week 12 and week 16 (Appendix C.11).

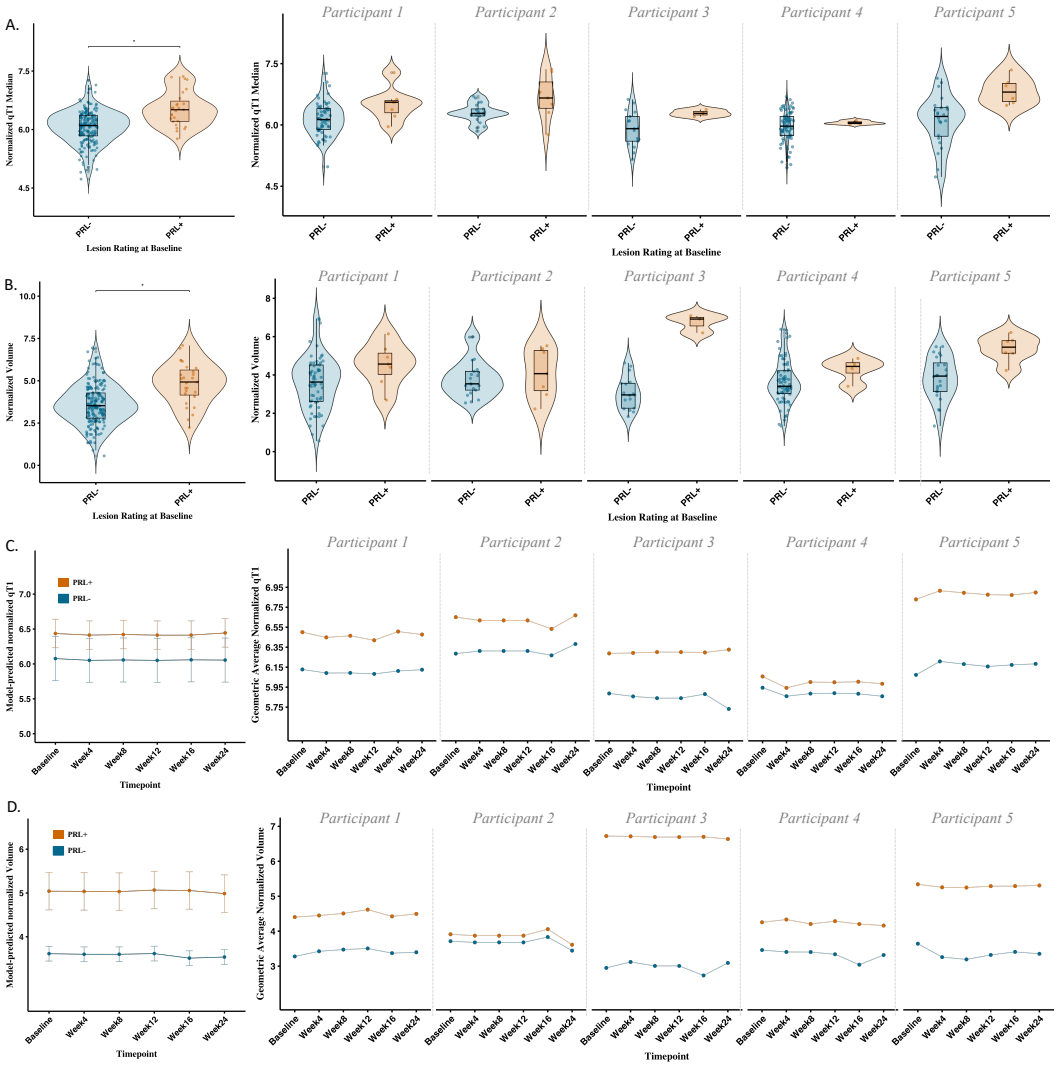


Figure 3.9: Lesion qT1 and volume in trial completers ($n = 5$). A) Normalized qT1 and B) normalized lesion volume was consistently higher in PRL+ compared to PRL- lesions, based on linear mixed models with participants as nested random effects. Right-side panels show each participant's average baseline values for PRL+ and PRL- lesions. C) Estimated normalized qT1 and D) normalized volume from the repeated-measures ANCOVA models comparing PRL+ and PRL- lesions over time. Left-side panels display participant-wise geometric averages for each timepoint.

3.2.3.4 Analysis of Clinical Outcomes

Changes in clinical disability scores across time are shown in Figure 3.10. Between baseline and week 24, participants showed small mean improvements in several clinical scores, including 9HPT dominant hand (-0.21 ± 0.60), 9HPT

non-dominant hand (-0.33 ± 1.30), PASAT ($+3.8 \pm 2.52$), SDMT ($+2.3 \pm 1.24$), and SNRS ($+2.0 \pm 1.6$). However, repeated one-way ANOVA analysis revealed no significant changes over time for any of these measures: 9HPT DH ($F(6,5) = 1.42$, $p = 0.30$), 9HPT nDH ($F(6,5) = 1.30$, $p = 0.32$), PASAT (Friedman statistic $= 5.34$, $p = 0.38$), SDMT ($F(6,5) = 2.85$, $p = 0.10$), and SNRS ($F(6,5) = 0.95$, $p = 0.43$). A linear mixed-effects model, accounting for repeated measures nested within participants, revealed a modest but statistically significant positive slope over time for both SNRS (0.08 ± 0.04, $p = 0.04$) and SDMT scores (0.05 ± 0.004, $p = 0.002$).

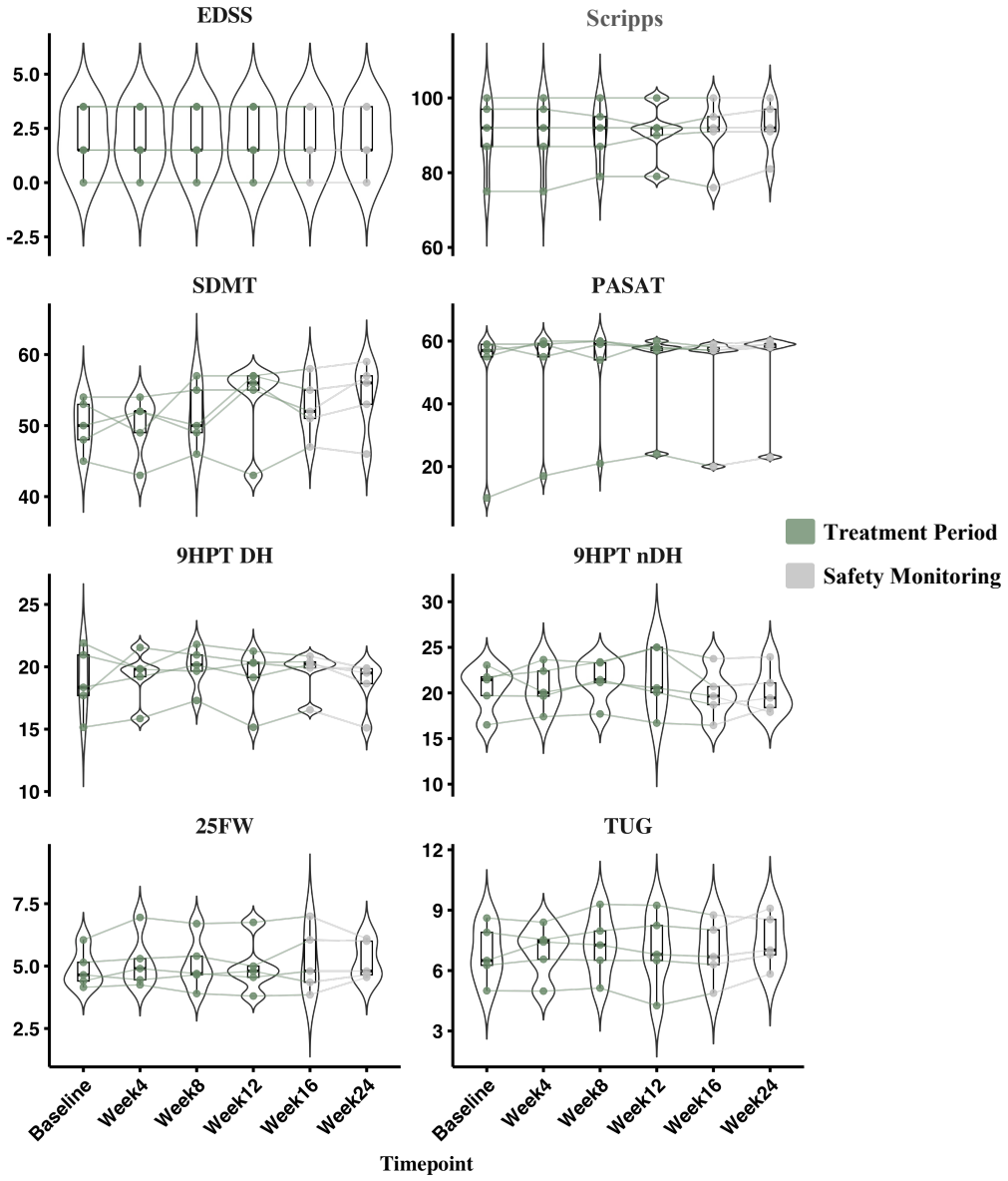


Figure 3.10: Longitudinal Changes in Clinical Scores. Violin plots showing changes in clinical scores from Week 0 to Week 24 for trial completers (n=5), with individual participant changes overlaid as line plots. Light green represents timepoints where participants were taking anakinra treatment, while grey represents time points in the safety monitoring period, after treatment had ended. Legend: EDSS = Expanded Disability Status Scale; SDMT= Symbol Digit Modalities Test; Scripps = Scripps Neurologic Rating Scale; PASAT = Paced Auditory Serial Addition Test; 9HPT DH= 9-Hole Peg Test for dominant hand; 9HPT nDH= 9-Hole Peg Test for non-dominant hand

3.2.4 Conclusions

As part of the ATaC-MS clinical trial, the potential of anakinra to modulate chronic inflammation in MS was evaluated by assessing changes in paramagnetic rim lesions. The primary objective was not met: no lesions demonstrated the disappearance of the paramagnetic rim following 12 weeks of treatment.

Secondary analyses revealed that both PRL+ and PRL- lesions remained stable in volume and qT1 values over time, with some evidence of PRL- decreasing in qT1 and volume within the trial. Although no changes were observed in PRL+ lesions over the course of the trial, prior studies, including the study from the previous section (Chapter 3.1), have demonstrated that these lesions typically expand slowly and are associated with ongoing tissue damage (Dal-Bianco et al., 2021; Elkady et al., 2025). Based on such studies, it would be reasonable to expect increases in lesional qT1 and volume in the absence of treatment. In this context, the observed lesion stability could suggest a treatment effect; however, in the absence of a control group and established PRL+ lesion progression, definitive conclusions cannot be drawn.

Clinically, modest improvements were observed in some clinical scores. Linear mixed effects showed slight but statistically significant improvements in Scripps and SDMT scores. However, the changes were small in magnitude and observed in a very limited participant sample ($n = 5$) and without a control group. Additionally, practice effects, particularly on cognitive measures, cannot be excluded. Taken together, these factors restrict the strength of conclusions drawn from the clinical score findings. Overall, while this study provides a structured framework for imaging-based monitoring of chronic lesion activity, more extensive, controlled trials are necessary to evaluate the therapeutic potential of anakinra in this context.

The imaging results and analysis for this trial will be discussed in Section 3.4 together with the results from the BRaKE-MS trial (Section 3.3).

3.3 Effect of Tolebrutinib on Chronically Inflamed White Matter Lesions in Multiple Sclerosis

In this section, I describe the 7T MRI evaluation of the investigational drug tolebrutinib on chronically inflamed WMLs in the BRaKE-MS trial. Following the structure of Chapter 3.2, I first provide background on tolebrutinib in MS, then describe and analyse the trial's 7T MRI WML endpoints. While the image processing methods are largely consistent with those in the previous section, this trial differs in design, including alternate follow-up timelines and a reference group; therefore, the methods are re-described as needed. This

study aims to assess whether tolebrutinib modulates paramagnetic rim status or alters qT1 and volume trajectories of PRL+ lesions relative to PRL– lesions.

3.3.1 Overview of Bruton's tyrosine kinase and Tolebrutinib in MS

Bruton's tyrosine kinase (BTK) is a non-receptor tyrosine kinase that plays a central role in immune cell signaling. It is expressed in hematopoietic cells, including B cells and myeloid cells such as microglia and macrophages. In B cells, BTK is activated when antigens bind to the B cell receptor (BCR), initiating a signaling cascade that involves the phosphorylation and activation of phospholipase C gamma 2 (PLCγ2). PLCγ2 hydrolyzes PIP2 (a membrane phospholipid) into IP3 and DAG, which act as second messengers. IP3 increases intracellular calcium, activating calcineurin, which in turn activates the nuclear factor of activated T cells (NFAT). DAG activates PKCβ, leading to the activation of NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells). Together, these transcription factors (NF-κB and NFAT) enter the nucleus and promote the transcription of genes involved in B cell activation, proliferation, and pro-inflammatory cytokine production (Alu et al., 2022). These cytokines can further promote the differentiation of pathogenic CD4+ T cells, contributing to demyelination in autoimmune diseases such as MS (Geladaris et al., 2022).

In myeloid cells (including microglia), BTK is activated downstream of Fcγ receptors, which are immune receptors engaged during antibody-mediated responses (Keaney et al., 2019). As in B cells, BTK activation leads to phosphorylation of PLCγ2 and downstream activation of the NFAT and NF-κB pathways. This results in transcriptional programs that promote microglial activation and a pro-inflammatory phenotype (Alu et al., 2022). Activated microglia are known to contribute to demyelination by releasing cytotoxic mediators and recruiting other immune cells.

In the context of MS, BTK's functions are highly relevant because pathogenic B cells and myeloid cells underlie both acute and chronic inflammation. Histological studies and transcriptomics analyses show that BTK is upregulated in microglia and macrophages at the rims of chronic active WML (Elkjaer et al., 2023). Elkjaer et al. (2023) found that BTK expression was dramatically enriched ($\sim$120-fold) in microglia at chronic active lesion edges, whereas other CNS cells were expressed far less, suggesting that elevated BTK in microglia

may be important in progressive aspects of MS. Thus, BTK sits at the nexus of the adaptive and innate drivers of MS: it mediates B cell pathogenicity via BCR signalling, and it transduces innate activation signals in CNS-resident cells. For these reasons, BTK has emerged as a promising therapeutic target in MS (Geladaris et al., 2022).

BTK inhibitors block BTK's kinase activity by occupying its ATP-binding pocket. Irreversible BTK inhibitors such as evobrutinib, tolebrutinib, remibrutinib, and orelabrutinib bind covalently to the cysteine-481 residue in BTK, permanently inactivating the enzyme. In contrast, reversible inhibitors (e.g., fenebrutinib) bind non-covalently within the kinase domain (Garcia-Merino, 2021). The inhibition of BTK interrupts the BCR-PLCγ2-NF-κB/NFAT cascade, thereby suppressing B cell activation, differentiation, and antibody production (Kramer et al., 2023). It also curtails downstream signalling from TLRs and Fcγ receptors in myeloid cells, limiting pro-inflammatory cytokine release. Thus, BTK inhibitors dampen both antigen-driven B cell responses and innate immune activation.

Preclinical studies support the therapeutic rationale of BTK inhibition in MS models. In EAE, oral BTK inhibitors attenuate disease mechanisms. Evobrutinib treatment reduced B cell activation and myeloid cell inflammation, shifting macrophages toward an anti-inflammatory phenotype (Geladaris et al., 2022). BTK inhibition has also shown tissue-reparative effects in demyelinating models. For example, evobrutinib markedly promoted remyelination in organotypic CNS slice cultures and even in a transgenic Xenopus tadpole model of demyelination (Martin et al., 2020). These models are independent of adaptive immunity, and immunostaining revealed that the cells expressing BTK in lesions were predominantly microglia, suggesting that inhibiting microglial BTK may create an environment favourable to repair. Thus, relevant proposed CNS effects of BTK inhibition include blocking microglial activation, promoting remyelination (via oligodendrocyte support), and preventing further immune-cell influx (Geladaris et al., 2022).

While no BTK inhibitors have yet been approved for MS treatment, several are currently being investigated in clinical trials (Airas et al., 2024). BTK inhibitors have garnered significant interest as DMTs for MS, owing to their therapeutic advantages and ability to target both adaptive and innate immune mechanisms. These small-molecule agents, which can be taken orally and have

intracellular access, offer a unique opportunity to modulate the immune response on both sides of the BBB. This dual action allows BTK inhibitors to target both peripheral immune cells, which are involved in MS relapses, and CNS-compartmentalized inflammation, thought to contribute to disability accumulation in progressive forms of MS. In relapsing forms of MS, BTK inhibition is hypothesized to reduce acute inflammation and modulate B lymphocytes, rather than depleting them, which may potentially reduce contrast-enhancing lesions. In progressive MS, inhibiting BTK within the CNS may have beneficial effects on chronic lesions and meningeal inflammatory infiltrates, making BTK inhibitors a promising therapeutic approach for both relapsing and progressive MS.

Tolebrutinib (formerly BIIB091) is a next-generation BTK inhibitor engineered for high potency and CNS penetration. It is an orally available, irreversible inhibitor that covalently binds BTK. Compared to other BTK inhibitors, tolebrutinib has been shown to achieve CSF concentrations higher than the necessary IC_{90} for BTK inhibition (Turner et al., 2024). In the same study, in-vitro kinase assays confirmed that tolebrutinib associates with BTK approximately 65-fold faster than evobrutinib. Thus, tolebrutinib combines high biochemical potency with robust CNS pharmacokinetics – a profile ideal for tackling CNS-intrinsic inflammation. By achieving effective BTK blockade in both peripheral B cells and CNS-resident cells (microglia, macrophages), tolebrutinib can, in principle, suppress the key drivers of both relapse activity and chronic progression.

Early clinical trials have demonstrated tolebrutinib's efficacy in MS. In a phase 2b study (Reich et al., 2021), relapsing MS patients receiving tolebrutinib (5–60 mg daily) showed a clear dose-dependent fall in new inflammatory lesions on MRI (Reich et al., 2021). At 12 weeks, the mean count of new gadolinium-enhancing lesions per patient was 0.13 in the 60 mg group, compared to 1.03 on placebo ($p = 0.03$). No new gadolinium lesions were observed in any patient on tolebrutinib after 12 weeks. Tolerability was good: only one serious adverse event (an MS relapse) occurred in the highest-dose arm, and no dose-limiting toxicities were seen in up to 120 mg single-dose exposure in phase I studies. These findings were sufficient to advance tolebrutinib into phase III trials. The phase III HERCULES trial (Fox et al., 2025) tested tolebrutinib in non-relapsing SPMS. Over a median of ∼133 weeks, significantly fewer patients on tolebrutinib reached confirmed 6-month disability progression than on

placebo (22.6% vs 30.7%; hazard ratio 0.69, 95% $CI = 0.55$–$0.88; P = 0.003$). This is the first demonstration that a BTK inhibitor can slow progression in SPMS. Adverse events were generally manageable, though a slightly higher rate of serious events (15.0% vs 10.4%) and grade $\geq$3 liver enzyme elevations were reported in the tolebrutinib group; most such elevations were reversible on discontinuation.

Chronic active MS lesions are characterized by a rim of activated, iron-laden microglia, which is visible on MRI as paramagnetic rim. Histologically, these lesions contain meningeal B cells and a shell of microglia and macrophages that sustain local tissue damage. Elevated BTK expression at lesion edges suggests BTK signalling plays a key role in maintaining this compartmentalized inflammation (Elkjaer et al., 2023). Unlike large antibodies or many disease-modifying therapies, tolebrutinib crosses the BBB and achieves high CNS exposure (Turner et al., 2024), enabling direct inhibition of BTK in both microglia and CNS-resident B cells. By dampening BTK-driven activation, tolebrutinib may disrupt the chronic inflammatory cycle at lesion rims. In the BRaKE-MS Phase 2 trial, the hypothesis that tolebrutinib reduces chronic lesion activity by inhibiting BTK-driven inflammation was tested, using PRLs as an imaging biomarker of treatment response.

3.3.2 Methods

3.3.2.1 Overview of Clinical Trial Design

This section uses data from an ongoing clinical trial titled *A Phase 2 Clinical Trial of Tolebrutinib, a Brain-Penetrant Bruton's Tyrosine Kinase Inhibitor, for the Modulation of Chronically Inflamed White Matter Lesions in Multiple Sclerosis* (BRaKe-MS, Appendix A.1). The trial is a single-site, phase 2, two-arm, rater-blinded but otherwise open-label study. Aspects of the clinical trial design and data collection relevant for this section are outlined in Figure 3.11. A total of 11 participants, who had been on stable anti-CD20 therapy for at least 6 months, enrolled in the trial. Participants were not randomized into the treatment/non-treatment cohorts, and participants who consented to be in the non-treatment cohort continued their anti-CD20 antibody therapy. In the treatment cohort, 60mg/day of tolebrutinib was initiated within three months of the most recent infusion of anti-CD20 monoclonal antibodies, and their anti-CD20 therapy was discontinued upon starting tolebrutinib. Four

participants (treatment cohort A) consented to escalate their tolebrutinib dose to 120mg/day after 96 weeks. However, due to a partial hold from the FDA, dose escalation for the remaining three participants (treatment cohort B) was cancelled, and they continued the 60mg daily dose of tolebrutinib for 96 weeks. Following screening, trial visits, including 7T MRI scans and clinical disease activity assessments, were scheduled every 12 weeks (±28 days for the non-treatment cohort and ±14 days for the treatment cohorts).

The clinical trial enrolment was restricted to adults with clinically progressive or stable MS, who had no clinical or radiological evidence of acute inflammatory activity within the past 6 months (i.e., no new WML or clinical relapses). All participants were required to have been on anti-CD20 therapy for at least 6 months and within 6 months of their most recent dose at the time of the baseline MRI scan. Additionally, participants were required to have at least one paramagnetic rim lesion visible on 7T MRI. The full inclusion and exclusion criteria for the trial are presented in Appendix C.3.2.

The MRI protocol was similar to the ATaC-MS trial, with the only difference in T2*-GRE acquisition (Appendix A.3.3). The same clinical tests used in the ATaC-MS trial were used to measure clinical disease activity: EDSS, 9HPT, 25FTW, Scripps, SDMT (see the last paragraph of Section 3.2.2.1)

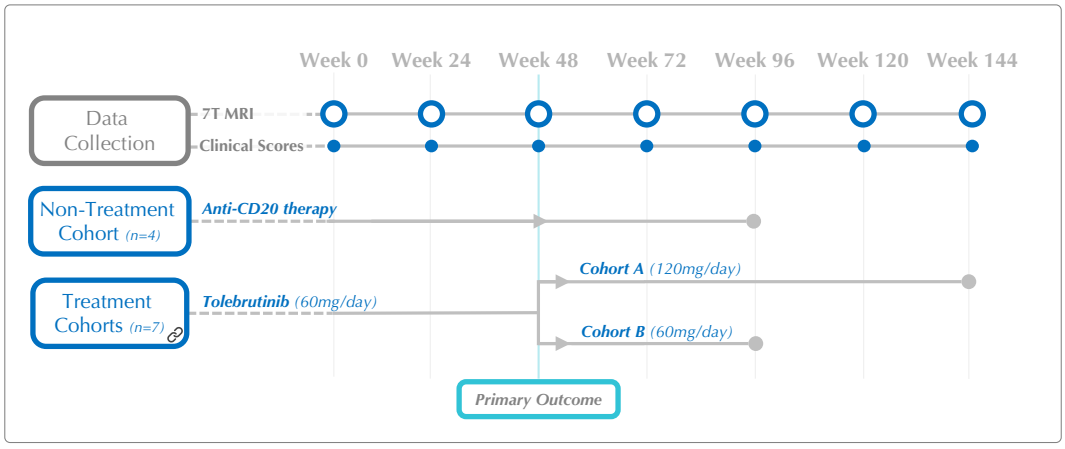


Figure 3.11: Overview of study design. 7T MRI and clinical disability scores were collected at baseline and every 12 weeks. Non-treatment participants were followed for 96 weeks. Treatment cohort was split into two sub-cohorts: cohort A in which participants escalate treatment dose at week 48 and were followed for 144 weeks and cohort B who continued original treatment dose for 96 weeks. Primary outcome endpoint was at week 48 (cyan blue).

3.3.2.2 Clinical Trial Endpoints

The primary outcome of this study was to evaluate the effects of 48 weeks of tolebrutinib 60 mg/day following B cell depletion on the paramagnetic rim of

chronically inflamed WML, as visualized on 7-tesla MRI. The primary endpoint was the per-patient proportion of lesions in which the paramagnetic rim had disappeared by the end of 48 weeks of daily 60mg tolebrutinib treatment. Secondary outcomes were longitudinal changes in qT1 and volume of PRL+ relative to PRL-. Exploratory tertiary endpoints included the proportion of paramagnetic rim lesions in which the rim had diminished or disappeared at any time point, changes in clinical scores, and any other MRI correlates indicative of reduced chronic neuroinflammation or tissue repair.

3.3.2.3 MRI Image Processing

MRI image processing followed the same steps as in the ATaC-MS trial (Section 3.2.2.3, Figure 3.6). Briefly, all MRI data were bias-field corrected, except for the estimated T1 maps. For PRL ratings, T2 * magnitude and phase slabs were stitched, intensity-normalized, and registered to upsampled baseline T1 maps (0.5 mm^3 isotropic). For quantitative analysis, native-resolution T1w images (0.7 mm isotropic) were segmented using PLAn 7T (Chapter 2.1). Baseline total lesion masks were split into individual lesions, manually corrected, and reviewed by a neurologist (M.I.G.). The corrected baseline lesion mask was then used in the boundary correction pipeline for longitudinal scans. PLAn 7T segmentations of NAWM were refined using the corrected lesion masks.

3.3.2.4 Radiological Assessment of the Paramagnetic Rim

Baseline Ratings of Paramagnetic Rim Status. Two neurologists (D.S.R. and M.I.G., again arbitrarily referred to as A and B from now on) separately rated all WML at baseline as having a paramagnetic rim (PRL+) or not having a paramagnetic rim (PRL-). This baseline rating was performed in the same way as described in the ATaC-MS trial (Section 3.2.2.4). Raters were not blinded to treatment/non-treatment cohorts. Any discrepancies in ratings between the two raters were discussed and resolved.

Longitudinal Ratings of Paramagnetic Rims Status. For the Primary Outcome Ratings, all lesions marked as PRL+ in the baseline ratings were longitudinally rated to see whether the rim itself changed over time. An older version of the Flask application (compared to the ATaC-MS trial, Section 3.2) was adapted for blinded longitudinal PRL ratings. Each webpage displayed participant-specific images of PRL+, each accompanied by a corresponding form. For every lesion, axial slices depicting the lesion were organized into

three anonymized columns representing distinct time points (Baseline, Week 24, Week 48). MRI contrasts for each lesion included estimated T1-maps, T2*-GRE magnitude, T2*-GRE phase echo 3, and T2*-GRE phase average registered to the participant's baseline T1w scan.

All participant IDs across cohorts were anonymized to allow the two raters to blindly evaluate changes in the paramagnetic rim with respect to the cohort. Longitudinal assessments were conducted between column pairs 1–2 (baseline - week 24) and 2–3 (week 24 - week 48). Rating options included: *rim more conspicuous*, *rim less conspicuous*, *no change*, *rim disappeared*, *rim appeared*, and *other*, with space for free-text comments. Here, each lesion was shown twice in chronological order. The order of lesions was randomly scrambled within each participant.

Lesions displaying inter- and intra-rater discrepancies were compiled for a consensus rating as previously described (Section 3.2.2.4).

3.3.2.5 Quantitative White Matter Lesion Analysis

As described in Section 3.2.2.5, median qT1 and volume were extracted for each lesion, the NAWM, and total lesion mask at all time points using native-resolution T1 maps (0.7 mm isotropic). Longitudinal changes in qT1 and volume were analysed for baseline-classified PRL+ and PRL- lesions within the T2*-GRE field of view. PRL+ lesions with rim status change were qualitatively compared to other lesions in the same participant to assess distinct qT1 or volume dynamics.

3.3.2.6 Statistical Analysis

Inter- and intra-reliability calculations of PRL ratings. Inter- and intra-rater reliability was assessed for baseline and longitudinal PRL ratings, pooling lesion ratings across all participants. Inter- and intra-rater reliability of PRL ratings were consistent with methods described in Section 3.2.2.6, using Cohen's unweighted (κ) for baseline ratings and weighted κ for longitudinal assessments. Importantly, unlike the longitudinal ratings in the ATaC-MS trial, lesions in this study were rated twice in chronological order, thereby eliminating the need to unscramble timepoints when assessing intra-rater reliability.

Primary Outcome Ratings. Both the total number of lesions exhibiting rim change across the Treatment and Non-treatment cohorts and the proportion of rim-changing lesions per participant were reported.

Quantitative Analysis of Lesion qT1 and Volume. As in the ATaC-MS trial (Chapter 3.2), the modelling framework captured both baseline and longitudinal changes in lesion qT1 and volume by PRL status. However, adjustments were necessary due to differences in trial design and participant characteristics.

To assess differences in qT1 and volume at baseline between PRL+ and PRL- lesions, a LMM was used, with a random intercept to account for participant-wise variations. To assess longitudinal changes in lesions, two types of LMMs were used: 1) a random coefficient model (RCM) with time as a continuous variable and 2) a repeated-measured mixed-effect (RMME) model with timepoint as a categorical variable to capture possible non-linear changes. To account for individual subject- and lesion-wise random variations, lesions were nested within participants as a random effect, with random intercepts and slopes. For RMME models, type III tests were used to evaluate fixed effects, and post hoc pairwise comparisons of estimated marginal means were performed within PRL+ and PRL- groups across timepoints, with Tukey-Kramer correction for multiple comparisons. Piecewise linear models were also explored but ultimately not adopted due to convergence issues, which stemmed from excessive breakpoint detection by the GLMMIX procedure.

To account for individual subject- and lesion-wise random variations, lesions were nested within participants as a random effect, with random intercepts and slopes. For RMME models, type III tests were used to evaluate fixed effects, and post hoc pairwise comparisons of estimated marginal means were performed within PRL+ and PRL- groups across timepoints, with Tukey-Kramer correction for multiple comparisons. Model diagnostics included assessing the normality of studentized and Pearson residuals using the Kolmogorov-Smirnov test and evaluating residual skewness and kurtosis. Statistical outliers were defined as observations with studentized residuals exceeding ± 3.

For all models, participant-wise covariates considered in the design of the statistical models included sex, baseline age, baseline normalized NAWM volume, baseline normalized NAWM median qT1, baseline normalized total lesion volume, and baseline normalized total lesion qT1 median. Lesion-wise covariates considered included baseline normalized lesion volume and baseline normalized qT1 median. Fixed effects were retained if they were not collinear and contributed significantly to the model, with significance defined as $p < 0.1$. The Treatment and non-treatment cohorts were analyzed separately due to

the non-randomized design. Additional RRME models were fit to treatment sub-cohorts.

All available longitudinal data were used (week 0-96 for the non-treatment cohort; week 0-96 or 0-144 for the treatment cohorts A and B, respectively). All median qT1 and volume values were normalized as described previously (Section 3.1.2.4). Some participants had extensive confluent lesions, skewing lesion volume and qT1 distributions. These outliers, unevenly distributed across participants, destabilized the statistical models and complicated the interpretation at the group level. All lesions manually identified as confluent at baseline were excluded to improve model convergence and reduce bias.

Changes in Clinical Scores. Clinical test scores were obtained at baseline and follow-up timepoints, provided participants were willing. For timepoints with missing scores, the missing value was imputed as the average of the scores from the previous and next timepoints. If the last timepoint was missing a score, the previous timepoint's result was carried forward. After testing for normality, changes in disability scores over time were analysed using one-way ANOVA for both treatment and non-treatment groups. Pairwise comparisons were performed between week 0 and week 48/week 96 using the Friedman test or the Holm-Sidak test for multiple comparisons with repeated measures. In the treatment cohort, changes between week 48 and week 96/week 144 were compared separately for cohort A (which escalated treatment at week 48) and cohort B.

3.3.3 Results

3.3.3.1 Participant Characteristics

This study included 11 participants: 7 in the treatment cohort and 4 in the non-treatment cohort. Table 4.6 summarizes baseline characteristics for each cohort. While most clinical and demographic variables did not differ significantly between cohorts, the treatment cohort had a notably higher total WML volume ($12,206 \pm 8802\ mm^3$) than the non-treatment cohort ($3,328 \pm 1,755\ mm^3$). Additionally, although the mean total lesion count was similar, the treatment cohort included participants with much higher lesion burdens, including a maximum PRL+ lesion count of over 40 lesions, greater than that in the non-treatment group.

Table 3.6: Participant demographics, whole brain MRI characteristics, lesion-level MRI characteristics, and clinical scores for Non-Treatment and Treatment Cohort. Data shown as mean ± SD, with ranges [min, max] where applicable.

		Non-Treatment (n=4)	**Treatment (n=7)**
Participant Demographics	Age (years)	40.5 ± 2.6	48.2 ± 3.0
	Sex (Female, Male)	3, 1	3, 4
	Sub-Diagnoses (RRMS, SPMS, PPMS)	2, 2, 0	2, 3, 2
	Race (White, Black/African)	4, 0	7, 0
	Ethnicity (Not H/Lat, H/Lat)	3, 1	7, 0
Whole Brain MRI Characteristics	NAWM Volume (mm^3)	457624 ± 21238	477553 ± 58418
	NAWM qT1 (ms)	1190 ± 18.6	1217 ± 45.6
	Total WML Volume (mm^3)	3228 ± 1755	12206 ± 8802
	Total WML qT1 (ms)	1651 ± 111	1805 ± 138
Individual Lesion MRI Characteristics	# WML	22.5 ± 25.7 [20, 77]	28.2 ± 28.5 [17, 83]
	# PRL+	5.2 ± 1.6 [3, 8]	8.7 ± 6.4 [1, 13]
	# PRL-	40 ± 26.7 [17, 73]	48 ± 28.6 [16, 71]
	WML qT1 (ms)	1627 ± 268	1697 ± 308
	PRL+ qT1 (ms)	1809 ± 320	1933 ± 336
	PRL- qT1 (ms)	1594 ± 244	1629 ± 263
	WML Volume (mm^3)	84.9 ± 113	168 ± 599
	PRL+ Volume (mm^3)	92.2 ± 65.6	306 ± 856
	PRL- Volume (mm^3)	83.5 ± 120	128 ± 495
Clinical Scores	EDSS	2.6 ± 2.8 [0, 6.5]	4 ± 1.7 [1.5, 6]
	Scripps	83.8 ± 17.1 [60, 100]	81.3 ± 7.1 [70, 90]
	SDMT	55.8 ± 12.3 [40, 70]	48.4 ± 7.7 [39, 63]
	PASAT	54.3 ± 4.9 [47, 57]	50.3 ± 8.9 [37, 60]
	9HPT DH	22.8 ± 6.9 [17.3, 32.9]	22.8 ± 3.0 [19.3, 27.1]

Abbreviations: NAWM = normal appearing white matter; qT1 = quantitative T1 times; WML = white matter lesion; PRL = paramagnetic rim lesion; RRMS = relapsing-remitting multiple sclerosis; SPMS = secondary progressive multiple sclerosis; PPMS = primary progressive multiple sclerosis; H/Lat = Hispanic/Latinx; EDSS = Expanded Disability Status Scale; Scripps = Scripps Neurological Rating Scale; SDMT = Symbol Digit Modalities Test; PASAT = Paced Auditory Serial Addition Test; 9HPT DH = 9-Hole Peg Test, dominant hand; 25 FTW = 25-Foot Timed Walk; TUG = Timed Up and Go. *One completer did not complete PASAT at baseline.

3.3.3.2 Radiological Assessment of the Paramagnetic Rim Analysis

Baseline inter-rater agreement for the classification of lesions as PRL+ or PRL- was substantial ($\kappa = 0.63$, Table 3.7), with a raw percent agreement of 94%. Longitudinal intra-rater agreement of changes in rim status ranged between substantial or almost perfect (Table 3.7), while inter-rater agreement of longitudinal changes was fair (quadratic weighted Cohen's $\kappa = 0.24 - 0.36$). Interestingly, reliability metrics of ratings for Column 1 – Column 2 on the Flask Application (*Change in paramagnetic rim from Week 0 – Week 24*) tended to be higher than from Column 2 - Column 3 (*Change in paramagnetic rim from Week 24 – Week 48*) (Table 3.7). An overview of the PRL+ lesions included in the various ratings and analyses is provided in the Appendix (Appendix C.3.2, Figure C.2).

Table 3.7: Inter- and Intra-Rater Reliability of Radiological Paramagnetic Rim Lesion Ratings.

Description	**Type of Reliability**	**Sample size***	**Cohen's Kappa details**	**Cohen's Kappa**
Baseline Ratings				
Classification of all lesion segmentations at baseline as PRL+ or PRL-.	Inter-rater	823	Unweighted, z=18	0.62
Longitudinal: Primary Outcome Rating (Completers only)				
Change in paramagnetic rim from Week 0 – Week 24 in PRL+ lesions only.	Rating 1: Inter-rater	116	Quadratic weighted, $z = 4.14$	0.36
	Rating 2: Inter-rater	114	Quadratic weighted, $z = 3.06$	0.28
	Rater A: Intra-rater	140	Quadratic weighted, $z = 9.34$	0.79
	Rater B: Intra-rater	114	Quadratic weighted, $z = 9.10$	0.84
Change in paramagnetic rim from Week 24 – Week 48 in PRL+ lesions only.	Rating 1: Inter-rater	115	Quadratic weighted, $z = 3.89$	0.30
	Rating 2: Inter-rater	114	Quadratic weighted, $z = 2.89$	0.24
	Rater A: Intra-rater	142	Quadratic weighted, $z = 7.65$	0.64
	Rater B: Intra-rater	112	Quadratic weighted, $z = 8.39$	0.77

Abbreviations: PRL+ = paramagnetic rim lesions
* Lesions marked as *other* in any rating were excluded from reliability analyses. These lesions were reviewed during consensus ratings.

Table 3.8 summarizes the longitudinal changes of paramagnetic rim status by Week 48 in lesions rated as PRL+ at baseline. Of the 122 lesions rated as PRL+ at baseline (104 in the treatment cohort and 18 in the non-treatment cohort), only one lesion's paramagnetic rim became less conspicuous in the treatment cohort (Figure 3.12A). No other rim change was detected in any participants in either treatment or non-treatment cohort. No participants in the treatment cohort showed the disappearance of PRLs, indicating that the primary outcome criterion (disappearance of at least one paramagnetic rim lesion in 2 out of the 7 participants) for rejecting the null hypothesis was not met.

Table 3.8: Primary Outcome Results. Overview of Rim Changes in PRL+ from Week 0 – Week 48 by Cohort and Participant

	Treatment							**Non-Treatment**			
Total # PRL+ Lesions	104							18			
Total # of *Rim Disappeared*	0							0			
Total # of *Rim Less Conspicuous*	1							0			
Participant	B001	B002	B003	B004	B005	B006	B007	C001	C003	C004	C005
Total # PRL+ Lesions	14	15	9	7	45	6	8	7	2	3	6
Total # of *Rim Disappeared*	0	0	0	0	0	0	0	0	0	0	0
Total # of *Rim Less Conspicuous*	0	1	0	0	0	0	0	0	0	0	0
Total # of *No Change*	14	15	9	7	45	6	7	7	2	3	6
Per participant proportion of rims disappeared	0/14	0/15	0/9	0/7	0/45	0/6	0/7	0/7	0/2	0/3	0/6

Abbreviations: PRL+ = paramagnetic rim lesions
Excludes WMLs that could not be rated due to motion artifacts or because they were not visible in the follow-up T2-GRE field of view.

Compared to other lesions in the same participant, the lesion whose paramagnetic rim became less conspicuous between Week 0 and Week 48 exhibited a qT1 and volume profile similar to the participant's average PRL- lesions (Figure 3.12B, Figure 3.12C). The lesion's normalized qT1 time at baseline was only slightly higher than the participant's average PRL- qT1 and much lower than the participant's average PRL+ (Figure 3.12B). Additionally, the lesion's qT1 showed a modest downward trend over time, except for a fluctuation at Week 120. The lesion was smaller than the participant's average PRL+ and showed a slight increase from baseline to Week 48, followed by a gradual

decrease through Week 144, similar to the average PRL- trajectory (Figure 3.12C).

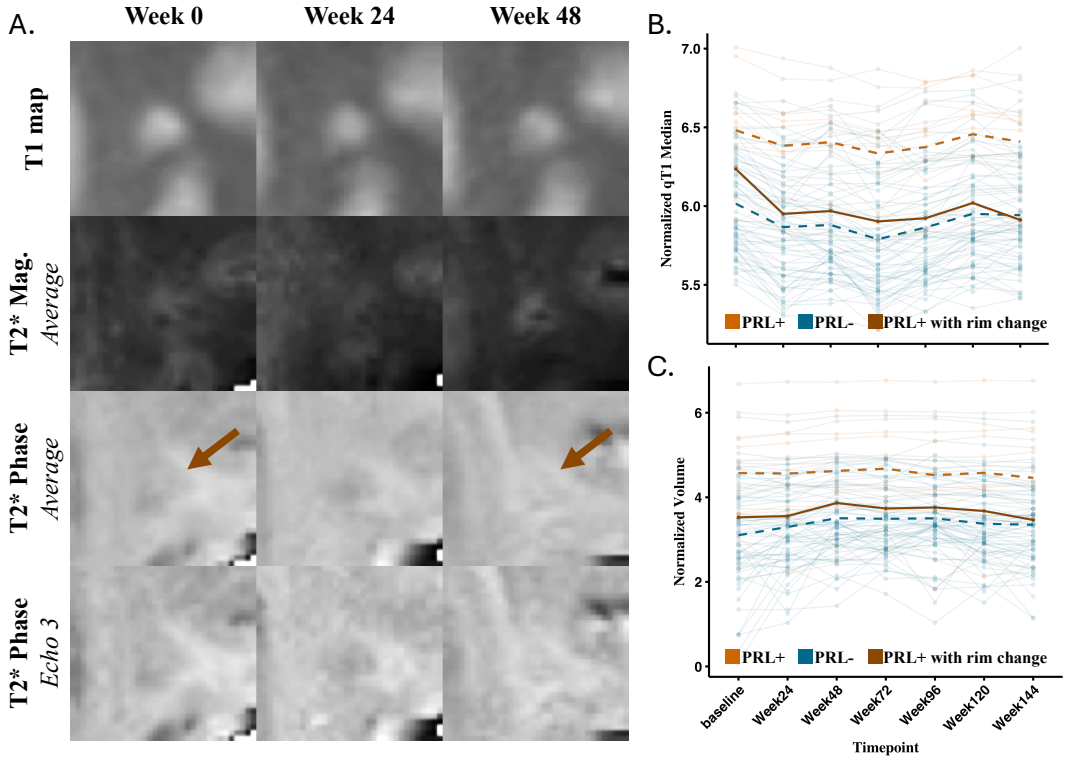


Figure 3.12: The PRL+ from a participant in treatment cohort A, with a rim that became less conspicuous over time. A) Longitudinal axial slices of T1 map, T2*-GRE magnitude, and phase images, all registered to Week 0, of the PRL+ lesion with a rim becoming less conspicuous from Week 0 (left column) to Week 48 (right column). B) For the participant with one PRL+ lesion that became less conspicuous over time, B) normalized qT1 median values and C) normalized Volume for all PRL+ and PRL- lesions are plotted over 144 weeks. The average values across PRL+ or PRL- lesions are shown with bold dashed lines. The solid dark orange line highlights the PRL+ lesion that became less conspicuous. (For details about quantitative analysis see Section 3.3.2.5

3.3.3.3 Quantitative White Matter Lesion Analysis

At baseline, PRL+ lesions exhibited significantly higher qT1 values in the treatment ($\beta_{\text{PRL+}} = 6.2 \pm 0.05$, $t(464) = 6.6$, $p < 0.001$) and non-treatment ($\beta_{\text{PRL+}} = 5.8 \pm 0.10$, $t(108) = 2.5$, $p = 0.01$) cohorts compared to PRL- lesions (treatment $\beta_{\text{PRL-}} = 5.9 \pm 0.09$; non-treatment $\beta_{\text{PRL-}} = 5.6 \pm 0.16$, Figure 3.13A) after adjusting for normalized lesion volume. Similarly, in both cohorts, PRL+ had higher lesion volume (treatment $\beta_{\text{PRL+}} = -3.4 \pm 0.1$, $t(415) = 5.4$, $p = 0.009$; non-treatment $\beta_{\text{PRL+}} = 4.1 \pm 0.3$, $df(111) = 2.4$) compared to PRL- (treatment $\beta_{\text{PRL-}} = -4.1 \pm 1.2$; non-treatment $\beta_{\text{PRL-}} = 3.5 \pm 0.1$, Figure 3.14B), after adjustment for normalized total baseline volume. Higher qT1 and larger volume in PRL+ lesions were qualitatively consistent across

participants, though the magnitude of the difference varied between individuals (Figure 3.13C, Figure 3.13D).

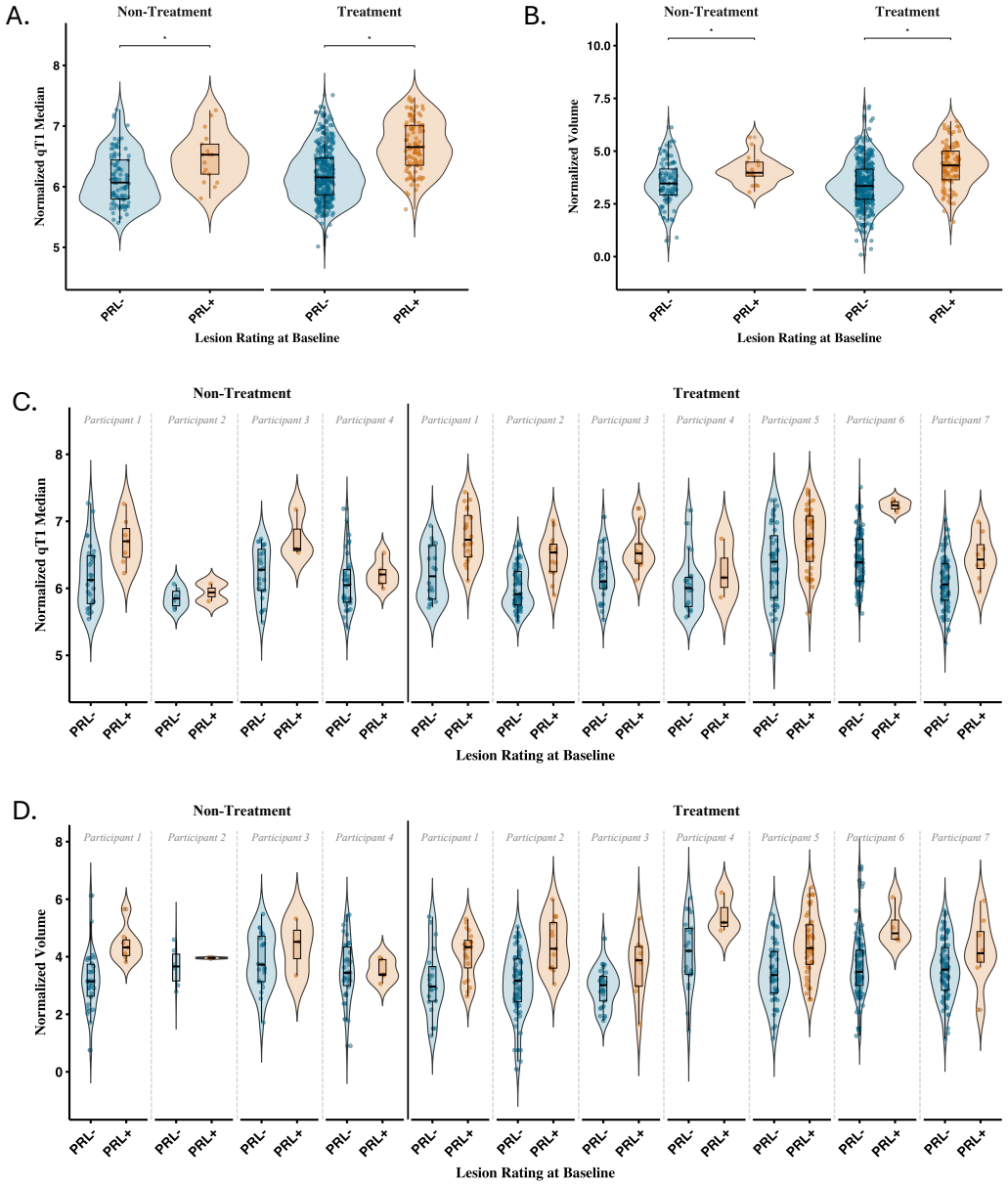


Figure 3.13: Baseline median qT1 and volume in PRL+ and PRL-. Normalized A) median qT1 and B) volume in PRL+ vs PRL- lesions in the Non-treatment and Treatment cohort. Statistical comparisons between PRL+ and PRL- lesions were performed separately for each cohort using linear mixed models with lesions nested within participant. Participant-wise normalized C) median qT1 and D) volume in PRL+ and PRL-. No statistical analysis was conducted on the participant-wise data due to insufficient lesion counts per group.

Figure 3.14 shows the longitudinal changes of qT1 and volume in PRL+ and PRL- lesions as assessed using the RCM. RCMs were fitted separately for the treatment and non-treatment cohorts, using all available longitudinal

data (weeks 0–96 for the non-treatment cohort and weeks 0–96 or 0–144 for the treatment cohort). The final qT1 models included covariates for baseline normalized total lesion volume, baseline normalized NAWM median qT1, and baseline normalized NAWM volume. The final volume models included covariates for baseline normalized total lesion volume only.

In the non-treatment cohort, PRL- lesions showed a significant decrease in qT1 over time ($\beta_{\text{PRL-}} = -4.7 \times 10^{-4} \pm 1.3 \times 10^{-4}$, $p < 0.001$), whereas PRL+ lesions did not ($\beta_{\text{PRL+}} = 1.6 \times 10^{-6} \pm 3.2 \times 10^{-4}$, $p = 0.99$, Figure 3.14A). In contrast, both PRL+ and PRL- lesions showed a significant decline in qT1 over time in the treatment cohort ($\beta_{\text{PRL-}} = -2.3 \times 10^{-4} \pm 5.9 \times 10^{-5}$, $p < 0.001$; $\beta_{\text{PRL+}} = -2.8 \times 10^{-4} \pm 1.1 \times 10^{-4}$, $p = 0.01$, Figure 3.14B). However, in both non-treatment and treatment cohorts, there was no significant interaction between PRL status and time ($F(1, 112) = 1.9$, $p = 0.2$; $F(1, 476) = 0.1$, $p = 0.7$), suggesting that the rate of qT1 change did not differ significantly between PRL+ and PRL-.

In both cohorts, the baseline normalized total lesion volume was significantly associated with normalized lesion qT1 (non-treatment: $p = 0.005$; treatment: $p < 0.001$), indicating that greater overall lesion burden corresponded to higher lesion qT1 values, independent of PRL status. In the treatment cohort only, baseline normalized NAWM qT1 was also positively associated with normalized lesion qT1 ($p = 0.04$), while baseline normalized NAWM volume showed no significant association with lesion qT1 in either cohort.

In models evaluating normalized lesion volume, lesion volume did not show a significant change over time in either PRL- ($\beta_{\text{PRL-}} = -9.0 \times 10^{-5} \pm 4.1 \times 10^{-4}$, $p = 0.8$) or PRL+ ($\beta_{\text{PRL+}} = 5.8 \times 10^{-4} \pm 1.0 \times 10^{-3}$, $p = 0.6$) lesions in the non-treatment cohort (Figure 3.14C). In the treatment cohort, PRL- lesions demonstrated a significant increase in volume over time ($\beta_{\text{PRL-}} = 7.8 \times 10^{-4} \pm 1.7 \times 10^{-4}$, $p < 0.001$), while PRL+ lesions showed no significant change ($\beta_{\text{PRL+}} = 1.2 \times 10^{-4} \pm 3.3 \times 10^{-4}$, $p = 0.7$, Figure 3.14D). However, the difference in slope between PRL- and PRL+ lesions was not statistically significant ($p = 0.08$). Baseline total lesion volume was a significant predictor of lesion volume in the treatment group ($p < 0.001$) but not in the reference group ($p = 0.8$)

The results of these linear mixed-effects RCMs indicated a modest fit to the data, with residual diagnostics revealing evidence of non-normality and some degree of skewness and kurtosis (data not shown). To address this, All RCMs were refit after excluding statistical outliers, which improved the model fit.

Importantly, the direction and statistical significance of the slopes remained consistent with the original results (Appendix C.3.1).

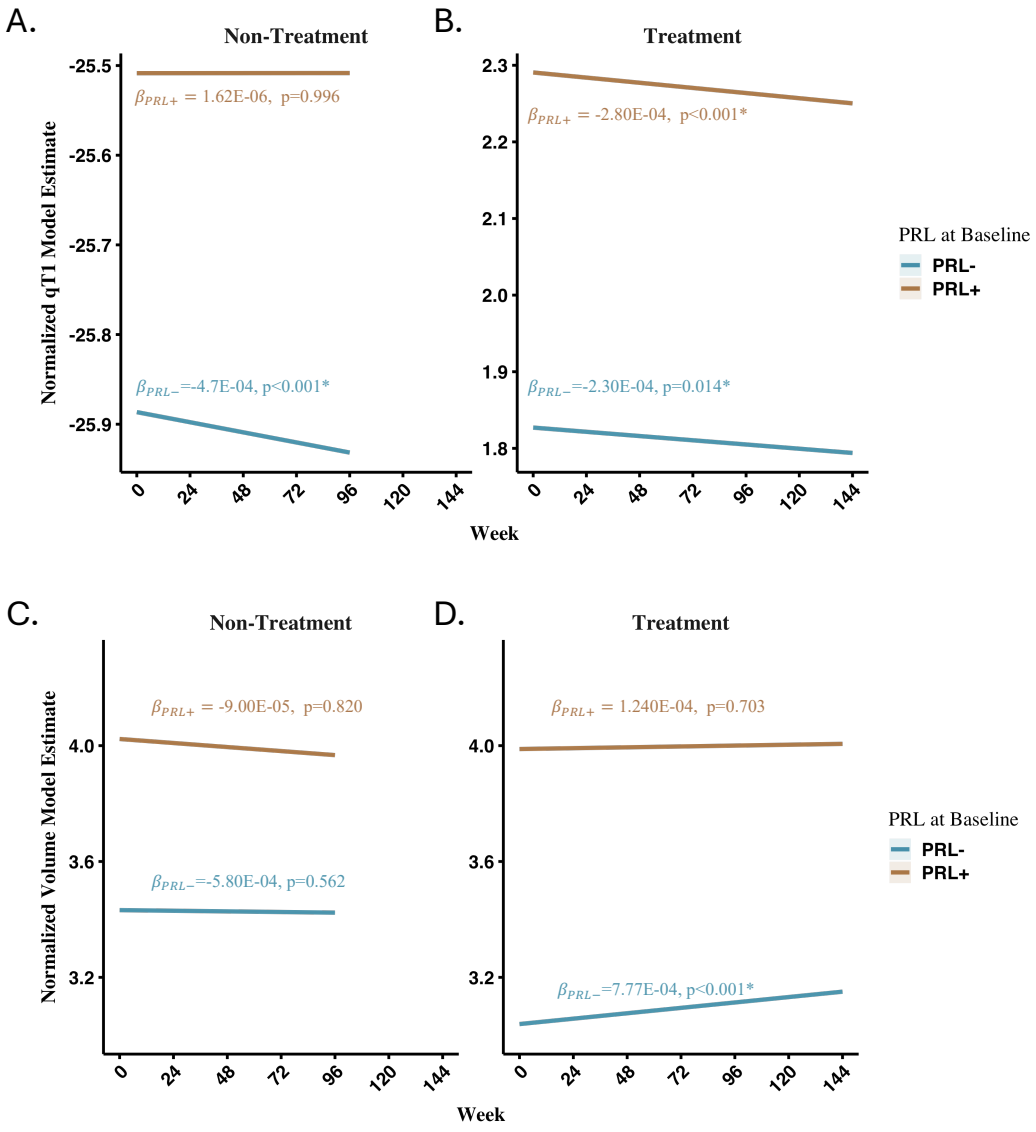


Figure 3.14: Longitudinal changes in qT1 and lesion volume in PRL+ and PRL- lesions, estimated using RCM. Predicted of slope estimates of lesion qT1 in PRL+ and PRL- in A) non-treatment and B) treatment cohort. Predicted of slope estimates of PRL+ and PRL- lesions volume over time in A) non-treatment and B) treatment cohort. All plotted lines are model estimates of normalized qT1 or volume. All available longitudinal data was used (week 0-96 for the non-treatment cohort; week 0-96 or 0-144 for the treatment cohort).
* Indicates if slopes are significantly different from 0. There was no significant difference between PRL+ and PRL- slopes within any of the models.

Additionally, to account for potential nonlinearity in the data, RMME models were fit with time points treated as categorical variables (Figure 3.15). All pairwise post-hoc comparisons are available in Appendix C.3.1.

In the non-treatment cohort, the interaction between PRL status and time was not significant ($F(4, 448) = 0.90$, $p = 0.5$), indicating that the time-specific changes of normalized lesion qT1 over time did not differ between PRL+ and PRL- lesions. No significant qT1 changes were observed within PRL+ lesions ($F(4, 448) = 0.4$, $p = 0.8$). Post hoc comparisons within PRL- lesions revealed that normalized lesion qT1 values significantly decreased from baseline to weeks 24, 48, and 72 (all $p_{\text{adj}} < 0.001$). The analysis controlled for baseline normalized lesion volume, which had a significant effect on normalized lesion qT1 ($F(1, 465) = 43.6$, $p < 0.001$).

In the treatment cohort, there was a significant effect of time on qT1 ($F(6, 2192) = 24.1$, $p < 0.001$), and the changes of normalized lesion qT1 across time points were different between PRL+ and PRL- ($F(6, 2193) = 5.8$, $p < 0.001$). Time also had a significant effect on normalized lesion volume ($F(6, 2192) = 2.3$, $p = 0.03$), but the interaction effect between PRL and time was not significant ($F(6, 2194) = 1.6$, $p = 0.1$).

Among PRL- lesions, normalized lesion qT1 decreased steadily in the early time points, with significant reductions observed from baseline to weeks 24, 48, and 72 (all $p_{\text{adj}} < 0.001$), but began to rise again at later time points. Post hoc comparisons between early and late weeks (e.g., week 72 vs. week 144) showed significant increases. Normalized qT1 in PRL+ lesions remained relatively stable over time, with fewer post hoc comparisons showing significant decreases. Notably, only one timepoint comparison (week 72 vs. week 144) revealed a significant increase ($p_{\text{adj}} = 0.03$).

In the treatment cohort, the volume trajectory of PRL- lesions mirrored that of qT1 (Figure 3.15A, right), with an initial increase in volume followed by a decrease after week 96. Post hoc tests revealed significant increases in lesion volume at weeks 48 and 72 compared to baseline, followed by a significant decrease at week 96 compared to baseline ($p_{\text{adj}} < 0.001$). PRL+ lesions did not significantly change over time ($F(6, 2193) = 0.22$, $p = 0.97$).

When the RMME model was fit separately to treatment cohorts A and B, PRL- lesions in treatment cohort A showed similar patterns to the overall treatment group: post hoc tests indicated an increase followed by a decrease in normalized lesion qT1, and a corresponding decrease followed by an increase in normalized lesion volume (Figure 3.15B, left). In treatment cohort A, PRL+ qT1 and volume were stable (Figure 3.15B, left). While there was a significant effect of time on qT1 ($F(6, 1052) = 2.7$, $p = 0.01$), post-hoc comparisons

showed no significant differences between timepoints. There was no effect of time on volume ($F(6, 1052) = 0.2$, $p = 0.99$).

In treatment cohort B, PRL- lesions showed significant post hoc decreases in normalized lesion qT1 between baseline and all follow-up timepoints ($p_{\text{adj}} < 0.001$, Figure 3.15B, right). PRL+ lesions also had significant post-hoc decreases between early time points (baseline-, week 48-) and week 72. In treatment cohort B, there was no significant effect of time or interaction between PRL and time on normalized lesion volume. There was no significant effect of time on normalized lesion volume ($F(4, 1132) = 1.04$, $p = 0.4$) and no significant interaction between PRL status and time ($F(4, 1132) = 0.1$, $p = 0.97$).

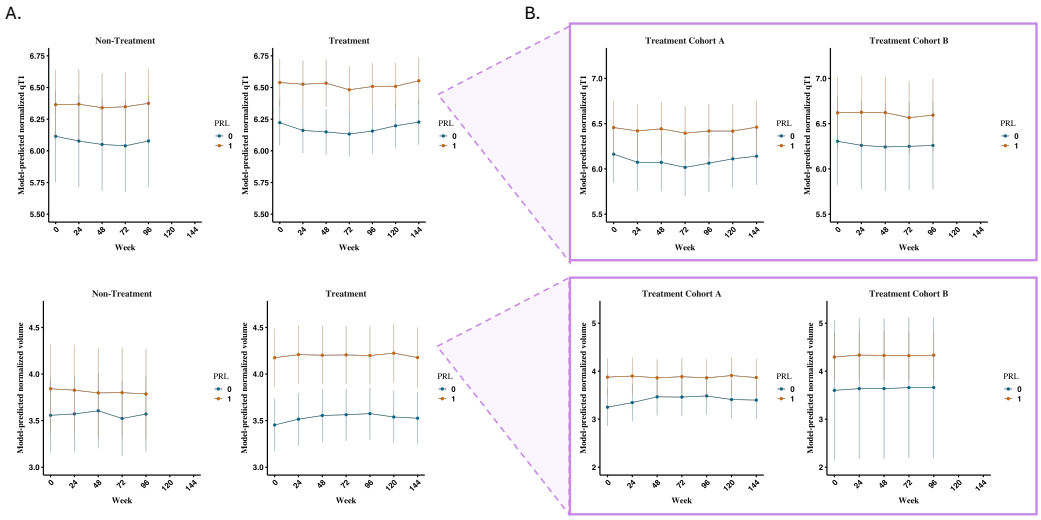


Figure 3.15: qT1 and volume trajectories in PRL+ and PRL- as measured by RRME model. A) Predicted estimates of normalized lesion qT1 and volume in PRL+ and PRL- lesions over time in non-treatment treatment cohorts. All plotted points are RRME estimates of normalized qT1 or volume ± SE. B) Predicted estimates of normalized lesion qT1 and volume in PRL+ and PRL- lesions over time in treatment cohort A and treatment cohort B. of See Appendix C.3.1 for pairwise post-hoc comparisons. All available longitudinal data was used (weeks 0-144 for treatment cohort A; weeks 0-96 for treatment cohort B).

3.3.3.4 Analysis of Clinical Outcomes

The only significant difference in clinical scores was a decrease in Scripps scores in the treatment cohort between week 0 (81.3±7.1) and week 96 (77.7±8; $t(6) = 3.12$, $p = 0.04$, Figure 3.16). This decrease was observed in both treatment cohort A and cohort B (Figure 3.16B), although repeated-measures ANOVA showed no significant change in Scripps scores over time within the subgroups (Cohort A: $F(1.31, 3.93) = 7.26$, $p = 0.053$; Cohort B: $F(1.068, 2.136) = 3.50$, $p = 0.20$), likely due to the small sample sizes (4 participants in Cohort A

and 3 in Cohort B). No prominent trends were observed between the non-treatment and treatment groups, with some participants showing steady disability scores, while others experienced worsening disability scores. Similarly, within the treatment cohorts, two participants in both Cohort A and Cohort B exhibited signs of worsening disability.

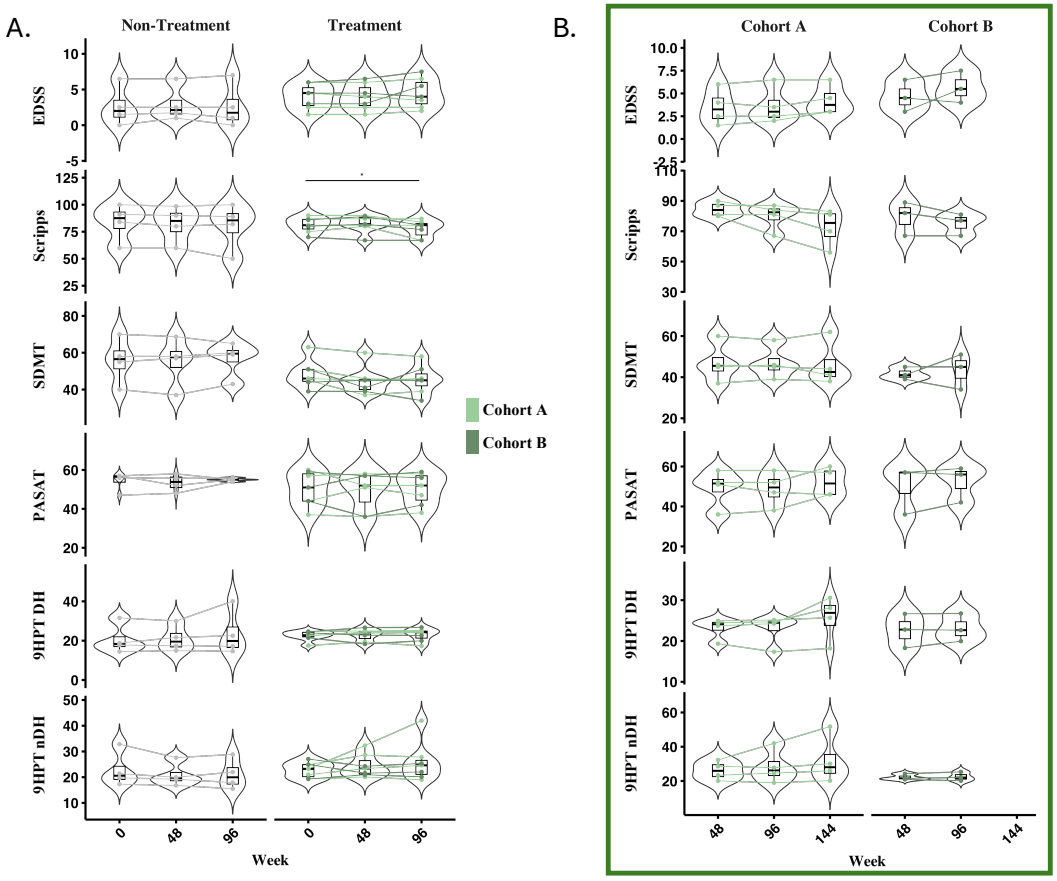


Figure 3.16: Longitudinal Changes in Clinical Scores Across Cohorts. A) Violin plots showing changes in clinical scores from week 0 to week 96 for non-treatment and treatment cohorts, with individual participant changes overlayed as line plots. B) Changes in clinical scores for treatment cohorts A and B at week 48, when cohort A escalated their tolebrutinib dose. Light green represents participants in treatment cohort A who escalated the dose at week 48 and were followed for 144 weeks, while dark green represents participants in cohort B who did not escalate the dose and were followed for 96 weeks. Significant differences in disability scores between timepoints were assessed using one-way ANOVA for each group, with pairwise comparisons between follow-up timepoints and week 0 (panel A) or week 48 (panel B). Legend: EDSS = Expanded Disability Status Scale; SDMT= Symbol Digit Modalities Test; Scripps = Scripps Neurologic Rating Scale; PASAT = Paced Auditory Serial Addition Test; 9HPT DH= 9-Hole Peg Test for dominant hand; 9HPT nDH= 9-Hole Peg Test for non-dominant hand

3.3.4 Conclusions

The BRAKE-MS trial aimed to evaluate whether 48 weeks of treatment with tolebrutinib, a brain-penetrant BTK inhibitor, could modulate the paramagnetic rim associated with chronically inflamed WML in MS. In this primary

outcome analysis of 104 lesions, no paramagnetic rims were resolved following treatment, suggesting that tolebrutinib did not reduce smoldering inflammation over the observed period sufficiently to modulate this imaging marker.

Several factors may explain the absence of rim change. First, the administered dose may have been insufficient to impact chronic inflammation. Although the study protocol included escalation to 120 mg for all seven participants in the treatment arm, only four received the higher dose before a clinical hold on all tolebrutinib trials was issued by the U.S. Food and Drug Administration due to reports of liver toxicity in unrelated trials. This deviation reduced the sample size and may have limited the statistical power to detect a treatment effect.

Second, it is possible that 48 weeks is not long enough to observe structural changes in PRLs. Ongoing longitudinal evaluations, including PRL ratings across all time points (up to week 144), will help assess whether a longer treatment of tolebrutinib modulates the paramagnetic rim. Furthermore, at the time of this writing participants are continuing tolebrutinib treatment in a long-term extension phase of the trial to assess the potential longitudinal treatment effect on chronic inflammation.

The lesion-level analysis of qT1 and lesion volume in the treatment cohort yielded mixed and ultimately inconclusive results. While RCM suggested a potential treatment effect in PRL+ lesions, evidenced by decreasing qT1 values over time with stable lesion volume, this pattern was not consistently supported by RMME modelling, which revealed an increase in qT1 between weeks 72 and 144.

PRL- lesions in the non-treatment cohort followed a trajectory of declining qT1 without accompanying volume change, consistent with less destructive pathology and potential for remyelination. In contrast, PRL- lesions in the treatment cohort showed a more complex profile, where qT1 initially declined but then increased after week 72. This fluctuation may reflect transient shifts in extracellular water content or be driven by limited sample sizes at later time points. As true remyelination within the first 72 weeks is unlikely, though not impossible, the biological relevance of this pattern remains uncertain. Importantly, in both RCM and RMME models, lesion volume mirrored the qT1 trajectory, (i.e., increasing volume when qT1 decreased, and vice versa) raising the possibility that the inclusion of surrounding tissue in the lesion segmentation may have influenced the measured trends. Whether these observed dy-

namics reflect underlying biology or methodological artifacts remains unclear. These findings, particularly regarding the role of image processing pipelines, are further explored in Section 3.4.

Participants treated with tolebrutinib did not experience clinical relapses or develop new MRI-detectable lesions over the 96-week period, consistent with low rates of lesion accumulation on tolebrutinib (Fox et al., 2025; Oh et al., 2025), especially in the context of the BRaKe-MS trial, in which lingering effects of B cell depletion may have contributed. The only clinical improvement observed was an increase in the Scripps Neurological Rating Scale score in the treatment cohort after 96 weeks, indicating better neurological function. All other disability measures remained stable across cohorts, suggesting no clear effect of tolebrutinib on overall clinical disability.

3.4 Discussion of MRI Results from Clinical Trials

The ATaC-MS (Chapter 2.2) and BRaKE-MS (Chapter 2.3) trials investigated whether pharmacological treatment could modulate chronic lesion inflammation, as measured by the potential disappearance of lesional paramagnetic rims. In addition to rim presence, both studies longitudinally tracked lesion volume and qT1 in PRL+ and PRL- lesions. The imaging analysis framework developed in Chapter 2.3.3 was successfully applied to both studies, despite differences in trial design and participant characteristics, demonstrating its overall suitability for clinical trials. This section will provide a combined discussion of the imaging findings from both trials.

3.4.1 Radiological Assessment of the Paramagnetic Rim

Lesions were categorized as PRL+ or PRL- at baseline, and then PRL+ lesions were longitudinally evaluated to determine changes in the paramagnetic rim intensity. Inter-rater reliability at baseline was substantial for both clinical trials, with Cohen's κ values around 0.6, falling within the average range reported in a literature review (Reeves, Mohebbi, Zivadinov, Bergsland, Dwyer, Salman, Schweser and Jakimovski, 2023). However, inter-rater reliability markedly declined for longitudinal PRL ratings, ranging from *no agreement* to *fair* agreement. These findings underscore the difficulty of qualitatively rating longitu-

dinal changes in rim intensity, a task that lacks well-established guidelines and has only recently begun to be explored in the literature (Absinta et al., 2021).

The intra-rater reliability of longitudinal changes in rim status differed between the two trials. The high intra-rater reliability in BRaKE-MS suggests that consistent interpretation of longitudinal changes in paramagnetic rim is achievable. Establishing a standardized training protocol for evaluating rim evolution could help improve inter-rater reliability. However, such efforts are currently limited by the small number of studies with sufficient longitudinal PRL data, making it challenging to develop robust training datasets. The intra-rater reliability in ATaC-MS was much lower than in BRaKe-MS, with one rater achieving $\kappa = 0.47$ and the other $\kappa = 0.3$, which was still higher than the inter-rater reliability for that trial. Several factors may have contributed to the reduced reliability observed in ATaC-MS, including differences in the T2*-GRE sequence used, a smaller number of lesions rated (25-40 compared to over 100 in BRaKE-MS), and a more randomized rating design where lesions were presented in either chronological or reverse-chronological order.

In both trials, some lesions initially classified as PRL+ at baseline were later deemed *not rateable* during longitudinal review and excluded from analysis. Although a binary baseline categorization of lesions as either PRL+ or PRL- was applied, an alternative approach could have involved rating lesions using a confidence scale, ranging from highly confident that a lesion is PRL+ to highly confident that it is PRL-. Using a graded scale at baseline to reflect the likelihood of PRL+ status could help reduce the number of lesions later reclassified as *not rateable*, particularly in cases where initial PRL determination was uncertain. Additionally, the confidence rating could be carried over in longitudinal assessments, where rim fading would correspond to a decrease in confidence in PRL+ classification.

The primary outcome was not achieved in either trial, with no PRL+ lesions resolving over the study periods. This may reflect the limited treatment durations, particularly in the ATaC-MS trial, where participants received the full dose for only four weeks. Given that prior research estimates the median lifespan of PRLs in untreated MS patients to be approximately seven years (Absinta et al., 2021), complete rim resolution may represent an ambitious target within the typical time frame of a clinical trial.

3.4.2 Quantitative White Matter Lesion Analysis

In both trials, individual lesion qT1 and volume over time in PRL+ and PRL- lesions were tracked. In addition, for the lesions that showed rim fading (one in each trial), qT1 and volume trajectories were visualized. In BRaKE-MS, the fading lesion resembled the baseline qT1 and size of PRL- lesions within the same participant and slightly decreased in qT1 over time (Chapter 3.3, Figure 3.12), supporting the idea that lesions with reduced rim intensity, reflecting reduced chronic inflammation at the lesion edge, may allow for possible remyelination. In contrast, the fading lesion in ATaC-MS retained high qT1 and volume relative to the rest of the participant's lesions, remaining more consistent with PRL+ profiles (Chapter 3.2, Figure 3.7). While rim fading is relatively infrequent in MS, future studies should explore whether specific lesion characteristics, such as size or baseline tissue integrity, predict the potential for rim resolution and remyelination.

In line with prior cross-sectional studies, PRL+ lesions in both trials were consistently larger and exhibited higher qT1 values than PRL- lesions, reflecting greater tissue damage (Absinta et al., 2019; Kolb et al., 2021). Longitudinal models revealed that PRL+ lesion volumes remained stable across all analysis models in both ATaC-MS and BRaKE-MS. At the same time, qT1 values were either stable or showed a modest decrease over time. These results could suggest a treatment-related effect, as PRL+ lesions are hypothesized to undergo gradual increases in qT1 in untreated MS. However, the interpretation of this potential treatment effect is limited by the observation that the non-treatment cohort in BRaKE-MS did not exhibit the expected increase in qT1 over time. In the natural history study (Chapter 3.1), PRL+ lesions showed significant increases in qT1 and volume over time; however, the cohort was heterogeneous, with participants receiving a variety of treatments. A separate 3T study (CONSONANCE trial) followed 95 participants treated with an anti-CD20 therapy, similar to those in the BRaKE-MS non-treatment cohort (Elkady et al., 2025). The study reported a decrease in tissue damage in PRL+ lesions over 96 weeks, while PRL- lesions showed the opposite trend. As anti-CD20 therapies do not readily cross the BBB, they are unlikely to directly target chronic inflammation in PRL+ lesions. This supports the hypothesis that tissue damage is worse in PRL+ lesions than in PRL-, and stabilization of qT1 in PRL+ could be a treatment effect of anakinra and tolebrutinib.

Several factors may explain why the BRaKE-MS non-treatment cohort did not exhibit a PRL+ trajectory as expected and as seen in the CONSONACE trial or natural history study in Chapter 3.1. The BRAKE-MS non-treatment cohort included only four participants and had considerable within-group variability. This arm comprised only four participants and exhibited lower baseline imaging markers of tissue damage (total lesion burden) than the BRaKE-MS treatment cohort. The imbalances between BRaKE-MS cohorts may reflect the open-label trial design, where individuals with milder disease may have been more likely to remain on their existing therapy rather than enrol in a clinical trial involving a novel treatment. While this underscores the value of having a well-matched reference group, identifying a "true control" in progressive MS is particularly challenging, as all participants are already at varying points within the disease course.

While PRL- lesions in the ATaC-MS trial remained relatively stable in both qT1 and volume, the BRaKE-MS trial showed a more complex behaviour of PRL-, with qT1 and volume initially decreasing before increasing again. Given that true remyelination is unlikely to occur within the first 72 weeks of observation, though not impossible, these fluctuations may instead reflect transient shifts in extracellular water content. Notably, qT1 and lesion volume appeared to mirror one another in the BRaKE-MS treatment cohort, suggesting that some of the observed changes may stem from image processing artifacts.

To further understand the potential influences of the image processing pipeline on the lesion qT1 and volume measurements, an exploratory analyses was conducted (Appendix C.3.2). Two key observations motivated this investigation. First, the consistent inverse relationship between lesion volume and qT1, particularly apparent in the PRL- of the BRaKE-MS treatment cohort, suggested that increased segmentation volume may artificially lower qT1 values by incorporating surrounding, less-damaged tissue. Notable fluctuations in both volume and qT1 across time points at the lesion, participant, and cohort levels in both clinical trials were observed. To explore these issues, I qualitatively examined (1) the impact of lesion segmentation on qT1 trajectories and (2) whether the affine registration was driving fluctuations in the longitudinal measurements. These exploratory analyses were conducted using the BRaKE-MS trial data, given the larger number of lesions and time points available compared to those in the ATaC-MS trial.

Using fixed lesion masks across timepoints, the characteristic qT1 trajectory in BRaKE-MS treatment cohort PRL- lesions – an initial decrease followed by a later increase – persisted (Appendix C.3.2, Appendix Figure C.4). Thus, the longitudinal qT1 changes in the BRaKE-MS treatment cohort A, as captured by the RRME model, are not primarily driven by segmentation variability; they either reflect true biological processes or simply capture statistically robust patterns that are not biologically meaningful due to the small, potentially noisy sample.

To assess whether registration influenced lesion volume and qT1 fluctuations, I extracted the Jacobian determinant (JD) from the affine transformation for each participant's follow-up timepoints. The JD, a measure that quantifies volume changes due to registration scaling, did not systematically explain lesion-level fluctuations (Appendix C.3.2, Appendix Figure C.5), suggesting that the observed trends are not primarily driven by registration-related artifacts.

While image processing may contribute to some variability in lesion measures, the overall trends captured in our qT1 and lesion analysis are not primarily driven by segmentation volume and registration artifacts.

3.4.3 Conclusions

In two clinical trials (ATaC-MS and BRaKE-MS), I explored whether therapeutic drugs (anakinra and tolebrutinib) could modulate chronic inflammation by assessing imaging biomarkers of inflammation and myelination. While limited, there was some indication of potential treatment-related effects. In two lesions, paramagnetic rims became less conspicuous over time. Additionally, the quantitative MRI analyses performed with the framework developed in this thesis revealed that PRL+ did not show progressive tissue damage, which was expected in untreated, chronically inflamed lesions. However, the absence of a true control group, small sample sizes, and inconclusive statistical findings preclude conclusions regarding therapeutic efficacy. Inclusion of additional timepoints and larger cohorts may be essential to reliably detect subtle therapeutic effects.

Additionally, in the quantitative analyses of longitudinal changes in lesional qT1 and lesion volume, lesion volume and qT1 values exhibited unanticipated behaviours, including abrupt fluctuations between time points. A series of

exploratory and sensitivity analyses were conducted to assess whether methodological factors, such as registration or segmentation, may have contributed to these patterns. No single methodological artefact emerged as a consistent or sufficient explanation for the observed variability, supporting the overall robustness of our longitudinal lesion analysis pipeline (Chapter 2.2).

Despite these limitations, the studies offer valuable insights into the potential use of paramagnetic rim disappearance as an imaging biomarker in clinical trials, an approach that has not been previously applied.

Chapter 4

Development of Ultra-Low Field MRI Analysis Tools for Tracking of Multiple Sclerosis Lesions

Contents

4.1 Adaptation of 7T PLAn for Ultra-Low Field Lesion Segmentation 152

4.2 Subtraction Imaging Pipeline to Assess Longitudinal Lesion Changes in Multiple Sclerosis at Ultra-Low-Field 182

4.3 Improving Ultra-Low-Field MRI Resolution using Partially Acquired K-Space 199

In this chapter, I aim to adapt, develop, and evaluate tools for segmenting, identifying, and tracking white matter lesions (WML) in multiple sclerosis (MS) using ultra-low field (ULF) MRI. High field (HF) MRI, typically between 1.5 and 3 tesla (T), is routinely used in MS to visualise and assess WML (Wattjes et al., 2021). However, the high cost and limited availability of these scanners can make regular monitoring difficult for patients (Geethanath and Vaughan, 2019). ULF MRI is a potential alternative that could help overcome these accessibility barriers (Wald et al., 2020). Unfortunately, the practical implementation of ULF MRI in the clinic is limited by low signal-to-noise ratio (SNR) and image quality (Arnold et al., 2023). Nonetheless, increasing numbers of studies have demonstrated the feasibility of ULF MRI in detecting WML, with one study showing that WMLs greater than 4 mm in diameter are radiologically detectable on ULF (Arnold et al., 2022). ULF MRI, though lower in image quality, may still have clinical relevance in settings such as triage, rou-

tine monitoring, and clinical trials. This chapter is divided into three sections, each addressing different aspects of WML analysis at ULF in MS.

In Section 4.1, I explore WML segmentation at ULF compared to HF and assess the performance of automatic segmentation algorithms at ULF. Within this context, I adapt our 7T segmentation tool, PLAn 7T (Chapter 2.1), to segment WML visible at ULF. All reference WML and demographic data for participants were created and extracted by Ashley Thomanaa (A.T.), and are indicated in the text with the initials A.T. Preliminary results from this section have been presented at a conference (Donnay et al., ISMRM BIC Symposium 2025, Thomanaa et al., ACTRIMS 2025).

In Section 4.2, I describe the development of a subtraction-based analysis pipeline to aid the radiological assessment of longitudinal WML changes. While previous studies have investigated the ability to radiologically detect MS lesions cross-sectionally on ULF MRI (Arnold et al., 2022; Okar et al., 2025), ULF MRI's capacity to support tracking lesion evolution remains unknown. I evaluate whether longitudinal lesion changes are detectable on ULF and assess whether a subtraction pipeline can aid these radiological assessments. Dr. Serhat V. Okar performed all radiological assessments, and these are indicated in the text with the initials S.V.O. The preliminary results from this section have been presented at a conference (Donnay et al., ISMRM 2024).

In Section 4.3, I aim to address a fundamental challenge of ULF MRI: low spatial resolution. In this section, I introduce a Fourier-based method that combines orthogonal anisotropic images to enhance ULF resolution. Preliminary findings from this work were presented at ISMRM 2024 and ECTRIMS 2023, and the final results have been published (Donnay et al., 2024).

4.1 Adaptation of 7T PLAn for Ultra-Low Field Lesion Segmentation

4.1.1 Introduction

Automated lesion segmentation of ULF MRI presents unique challenges. ULF's decreased SNR, reduced contrast, and lower spatial resolution compared to HF MRI make precise anatomical delineation challenging, which hinders accurate segmentation of disease-specific hallmarks, such as WML in MS.

While many automated segmentation tools exist for WML quantification on conventional MRI scanners (Griffanti et al., 2016; Selvaganesan et al., 2019; Valcarcel et al., 2018), segmentation methods for ULF MRI remain limited and show relatively low performance. The heterogeneous magnetic field properties, lower SNR, and lower resolution at ULF introduce artifacts similar to WML that complicate WML detection, raising questions about the generalizability of HF segmentation models to ULF scans (Arnold et al., 2022; Iglesias et al., 2023b). For instance, the method for inter-modal segmentation analysis (MIMoSA), a 3T WML segmentation algorithm that uses intermodal coupling regression and structural modelling, has demonstrated strong performance on 3T MRI (Valcarcel et al., 2018). However, in an initial comparison on 36 same-day paired 3T and 64mT data, MIMoSA's WML segmentation overlap between field strengths was poor, with a Dice Similarity Coefficient (DSC) of only 0.23, and some participants showed no overlap at all (Arnold et al., 2022). The low overlap may be due to limitations of applying HF-trained segmentation algorithms directly to ULF data, where artifacts like peripheral distortion and venous hyperintensities can lead to a high false positive rate. However, it may also be due to more fundamental limitations of comparing segmentations across such different field strengths. Using 3T segmentations as ground truth may not be a fair benchmark, as ULF's larger and anisotropic voxel sizes inherently cannot resolve small lesions visible on HF. Additionally, such a comparison would require registration between the two field strengths, which also introduces errors in the comparison, particularly because non-linear geometric distortions complicate the mapping between these field strengths. Thus, to properly assess whether HF algorithms such as MIMoSA generalise to ULF, evaluations should include a ULF-specific gold standard segmentation that reflects the realistic detection limits of ULF.

One approach to addressing segmentation challenges at ULF is to improve the quality of the ULF scans. *SynthSR*, a super-resolution deep learning (DL) model, synthesizes high-resolution isotropic T1-weighted (T1w) images from low-resolution data, making them compatible with existing HF neuroimaging tools, like segmentation algorithms (Billot et al., 2023; Iglesias et al., 2021). *SynthSeg*, a convolutional neural network (CNN) that uses domain randomisation, builds upon this approach by training on fully randomised synthetic data that varies in orientation, contrast, resolution, and artifacts (Billot et al.,

2023). This allows *SynthSeg* to handle a wide range of real MRI scans without requiring retraining or fine-tuning. Based on *SynthSeg*, WMH-SynthSeg is also designed to segment white matter hyperintensities (WMH) and, unlike the original *SynthSeg*, WMH-SynthSeg is tailored to handle ULF MRI. During training, WMH-SynthSeg incorporates data augmentation, mimicking resolution, reduced contrast, and partial volume effects typical in ULF scans. While the training cohorts include participants with vascular WMH and MS lesions, the performance evaluation was limited to volumetric comparisons between WMH-SynthSeg and 3T reference segmentations (Iglesias et al., 2023a,b). Additionally, WMH-SynthSeg has only been evaluated on a small cohort ($n = 12$) of MS participants (Iglesias et al., 2023b).

DL-based segmentation models, such as nnU-Net, provide another promising avenue for automating lesion segmentation at ULF. The nnU-Net framework self-configures its model architecture and preprocessing pipeline, and performs well in various 3D biomedical image segmentation tasks, including brain and WML segmentation in MS (Donnay et al., 2023; Isensee et al., 2021). Furthermore, transfer learning can improve WML segmentation performance when data availability is limited by reusing an already trained model as a starting point for a new, related task. In Chapter 2.1, I demonstrated that PLAn 7T, a transfer learning nnU-Net model pre-trained using readily available, 3T automatic segmentation labels and fine-tuned with 7T manual labels, had improved WML segmentation performance (Donnay et al., 2023). I hypothesised that adapting PLAn 7T to a 3T–ULF transfer learning algorithm could improve WML segmentation at ULF.

In this section, I first compare paired 3T and ULF scans to determine which lesions are visible and can be segmented on ULF images. Secondly, within the context of what WML segmentation is achievable at ULF compared to 3T, the performance of four automatic WML segmentation algorithms are evaluated on ULF data. For methods that allow the option of using more than one modality, the effect of using a T1w scan in addition to FLAIR in the model is also tested.

4.1.2 Methods

This retrospective study included 104 participants with known or suspected MS who underwent paired 64mT and 3T MRI scans on the same day between April 2021 and June 2024 (Appendix A.1, Natural History IRB protocol; Appendix A.4.1, MRI acquisition details).

4.1.2.1 Datasets for Training and Evaluation

All MRI acquisition details specific to the training and evaluation datasets are provided in Appendix A.4.1. Reference WML segmentations were created for all participants at both magnetic field strengths. At 3T, reference WML masks were generated by editing the WML output class of the Classification using Derived-based Features (C-DEF) segmentation algorithm, an in-house automated segmentation method (Selvaganesan et al., 2019). S.V.O. visually classified each participant's WML burden into one of three categories (mild, moderate, or severe), based on their 3T FLAIR scan. Mild cases had fewer than 10 WMLs. Moderate cases had between 10 and 20 WMLs and/or less than 25% of the lesion burden appearing confluent. Severe cases had 20 or more WMLs and/or more than 25% confluent lesion burden (Okar et al., 2025).

Using a consensus approach, two raters (A.A.T. and S.V.O.) manually annotated WMLs visible on ULF FLAIR images. Each hyperintensity was confirmed as a true-positive lesion by referencing a rigidly registered 3T FLAIR image. These manually annotated lesion masks are referred to as *ULF Aided*. To reproduce a more real-world scenario where a 3T scan is unavailable for reference, in a subset of 35 cases, an additional set of ULF WML masks was manually annotated, this time blinded to the 3T images, and these WML masks will be referred to as *ULF Blinded*.

The paired 3T and 64mT MRI scans and reference WML masks from both field strengths were split into a training ($n = 20$) and validation dataset ($n = 84$). The training dataset was created by randomly selecting seven mild, seven moderate, and six severe WML cases to ensure a diverse range of lesion burdens. Table 4.1 provides an overview of participant characteristics in both cohorts. Age, sex, and diagnoses across the training (mean age ± SD = 54 ± 2.6; female = 16) and validation cohorts (mean age ± SE = 48 ± 1.4; female = 62) were not significantly different ($p = 0.06$, $U = 609$; $p = 0.78$; $p > 0.99$; $p = 0.33$, $U = 695$). Additionally, WML volumes calculated from ULF Aided and 3T reference masks were not significantly different between training and validation cohorts ($p = 0.24$, $U = 696$; $p = 0.06$, $U = 613$).

All 64mT and 3T MRI scans were N4-bias field corrected and skull-stripped using *SynthStrip* (Hoopes et al., 2022; Tustison et al., 2010). The skull-stripped ULF T1w images were rigidly registered to the ULF FLAIR with linear interpolation using the ANTs toolbox (Avants et al., 2011). The resulting registration matrix was then applied to the non-skull-stripped images.

Table 4.1: Participant Characteristics and Descriptive Statistics in Training and Validation Cohorts. Data shown as count or mean ± SE, with ranges [min, max] where applicable.

		All Participants	**Training Cohort**	**Validation Cohort**
Participant Characteristics	# of participants	104	20	84
	Age in years	49.2 ± 1.3	54.3 ± 2.6	48.0 ± 1.4
	Sex (Female, Male)	78, 26	16, 4	62, 22
	Diagnoses (MS, OIND, NIND)	98, 3, 3	20, 0, 0	78, 3, 3
	Sub-Diagnoses (RRMS, PPMS, SPMS, RIS)	72, 7, 14, 5	12, 3, 3, 2	60, 4, 11, 3
	Race (White, Black, Multiracial, Asian, undisclosed)	80, 17, 3, 2, 2	14, 6, 0, 0, 0	66, 11, 3, 2, 2
MRI Characteristics	64mT WML volume (mL)	6.6 ± 1.0 [0, 70.6]	14.4 ± 4.2 [0, 70.6]	8.1 ± 1.1 [0, 49.6]
	3T WML volume (mL)	7.2 ± 1.0 [0, 46.8]	11.1 ± 2.9 [0.5, 46.8]	6.3 ± 1.0 [0, 43.7]

Abbreviations: SE = standard error; MS = multiple sclerosis; OIND = other inflammatory neurological disease; NIND = non-inflammatory neurological disease; RRMS = relapsing remitting MS; SPMS = secondary progressive MS; PPMS = primary progressive MS; RIS = Radiologically Isolated Syndrome; WML = white matter lesion; There are no significant differences between training and validation cohort characteristics.

4.1.2.2 Reference Segmentations

The overlap of 64mT and 3T reference WML segmentation masks were compared. For this analysis, the masks needed to be in the same coordinate space. ULF FLAIR scans and manual masks were up-sampled to match the voxel size of the 3T scans ($1mm$ isotropic). The skull-stripped 3T FLAIR scans were rigidly registered to the up-sampled skull-stripped ULF FLAIR using *FLIRT*, and the resulting registration matrix was applied to the 3T manual lesion masks (Jenkinson and Smith, 2001). To minimise resampling errors and preserve the original lesion volume, all resampled WML masks were linearly interpolated and thresholded at a value that minimised the difference in volume relative to the original lesion mask volume, selected using iterative thresholds in steps of 0.05.

Table 4.2: Overview of Automatic Segmentation Methods

Algorithm	**Method**	**Training Dataset**	**Input Contrasts**	**Output**
MIMoSA	Voxel-based weighted linear regression	3T FLAIR & T1w as described in original publication	64mT FLAIR, reg. 64mT T1w	WML probability map
SynthSeg	CNN (U-Net) trained with synthesized MRI	Trained with synthetic MRI as described in original publication	64mT FLAIR	Whole brain segmentation including a WMH label, in 1x1x1 mm^3 resolution
nnU-Net FLAIR	CNN (U-Net) with 5-fold CV	ULF FLAIR and ULF Aided WML masks from training dataset (n=20)	64mT FLAIR	Default binarised WML mask, SoftMax Probabilities
nnU-Net FLAIR/T1w	CNN (U-Net) with 5-fold CV	ULF FLAIR, reg. ULF T1w, and ULF Aided WML masks from training dataset (n=20)	64mT FLAIR, reg. 64mT T1w	Default binarized WML mask, SoftMax Probabilities
PLAn 1 FLAIR	CNN (U-Net) with 5-fold CV and transfer learning	Pre-trained down-sampled 3T FLAIR and C-DEF WML masks from original publication (n=20). Fine-tuned with ULF FLAIR and ULF Aided WML masks from training dataset (n=20)	64mT FLAIR	Default binarised WML mask, SoftMax Probabilities
PLAn 1 FLAIR/T1w	CNN (U-Net) with 5-fold CV and transfer learning	Pre-trained down-sampled 3T FLAIR, reg. 3T T1w and C-DEF WML masks from original publication (n=20). Fine-tuned with ULF FLAIR, reg. ULF T1w, and ULF Aided WML masks from training dataset (n=20)	64mT FLAIR, reg. 64mT T1w	Default binarised WML mask, SoftMax Probabilities
PLAn 2 FLAIR	CNN (U-Net) with 5-fold CV and transfer learning	Paired 3T and ULF training dataset: pre-trained down-sampled 3T FLAIR and manual 3T WML masks (n=20). Fine-tuned with ULF FLAIR and ULF Aided WML masks (n=20)	64mT FLAIR	Default binarised WML mask, SoftMax Probabilities
PLAn 2 FLAIR/T1w	CNN (U-Net) with 5-fold CV and transfer learning	Paired 3T and ULF training dataset: pre-trained down-sampled 3T FLAIR, reg. 3T T1w, and manual 3T WML masks (n=20). Fine-tuned with ULF FLAIR, reg. ULF T1w, and ULF Aided WML masks (n=20)	64mT FLAIR, reg. 64mT T1w	Default binarised WML mask, SoftMax Probabilities

Abbreviations: MIMoSA = Method for Inter-Modal Segmentation Analysis; CNN = Convolutional Neural Network; U-Net = U-shaped neural network; nnU-Net = no-new U-Net framework; CV = cross-validation; FLAIR = Fluid Attenuated Inversion Recovery; T1w = T1-weighted MRI; ULF = ultra-low field; WML = white matter lesion; WMH = white matter hyperintensity; PLAn = Pseudo-Label Assisted nnU-Net; reg. = registered; C-DEF = Classification using Derivative-based Features.

4.1.2.3 Overview of Automated Ultra-Low Field Segmentation Approaches

Automatic lesion segmentation algorithms were applied to the validation cohort of 64mT scans: MIMoSA, SynthSeg, nnU-Net, and PLAn, as outlined in Table 4.2.

MIMoSA: To create 64mT MIMoSA WML segmentations, skull-stripped and co-registered T1w and FLAIR images were normalised using the *WhiteStripe* method (Shinohara et al., 2014). These images were then segmented with a MIMoSA model previously trained on 3T FLAIR and T1w, as described in the original publication (Valcarcel et al., 2018). Based on the original publication, MIMoSA's output WML probability maps were thresholded at 0.2 and binarised. MIMoSA was included in the analysis to compare and build on existing findings in the field (Arnold et al., 2022).

SynthSeg: 64mT SynthSeg WML segmentations were obtained with FreeSurfer's WMH-SynthSeg tool using the preprocessed 64mT FLAIR images as input (Iglesias et al., 2023b). This tool outputs anatomical segmentation labels, including WML, at 1 mm^3 isotropic resolution, regardless of the input resolution. The WML labels were extracted from the output segmentation and binarised.

nnU-Net Models: Two nnU-Net segmentation models were trained using the publicly available nnU-Net package (https://github.com/MIC-DKFZ/nnUNet) and the ULF training dataset ($n = 20$) (Isensee et al., 2021). The first nnU-Net model was trained using only the preprocessed ULF FLAIR (nnU-Net FLAIR), and the second using the preprocessed ULF FLAIR and the preprocessed registered ULF T1w (nnU-Net FLAIR/T1w). Both models were trained using full 5-fold cross-validation and the *3d_cascade_fullres* configuration.

PLAn Models: The previously published transfer learning segmentation algorithm, PLAn-7T, was adapted to be compatible with ULF data. Four distinct models were trained: PLAn 1 FLAIR, PLAn 1 FLAIR/T1w, PLAn 2 FLAIR, and PLAn 2 FLAIR/T1w. These models were trained using either FLAIR images alone or a combination of FLAIR and registered T1w images during both pre-training and fine-tuning stages. In all models, transfer learning was performed as described in the original publication, by loading the weights of the pre-trained models, along with the corresponding processing parameters, excluding the final softmax layer (Donnay et al., 2023). The models were then fine-tuned using the ULF training dataset.

The first two models (PLAn 1 FLAIR and PLAn 1 FLAIR/T1w) were pre-trained using the dataset of 20 participants from the original publication, which included 3T FLAIR only or the 3T FLAIR and corresponding registered 3T T1w images, and C-DEF WML labels down-sampled to a resolution of 1.7×1.7×5 mm. These models were fine-tuned using only the preprocessed ULF FLAIR or the preprocessed ULF FLAIR and registered ULF T1w from the ULF training dataset, as described earlier. The second pair of PLAn models (PLAn 2 FLAIR and PLAn 2 FLAIR/T1w) was pre-trained and fine-tuned using the 3T-ULF paired training dataset. Pre-training was performed using down-sampled 3T scans and manual WML lesion masks (3T FLAIR only, or both 3T FLAIR and registered T1w images). The models were fine-tuned using the same participants' ULF scans.

4.1.2.4 Evaluation of Reference Segmentations

For both Aided ULF WML masks and the subset of participants with Blinded ULF WML masks, ULF WML volume and count were compared to the reference 3T WML masks. Lesion volume and count were calculated using scans in their the original resolution. Individual lesions were identified using connected-component analysis, where spatial clusters are defined by voxels connected by a shared face (FSL's *cluster* tool, connectivity=6) (Jenkinson et al., 2012). The total lesion count was calculated as the number of clusters larger than 10 mm^3. The DSC overlap between up-sampled manual ULF WML masks and manual 3T WML masks was assessed.

A total of 126 individual lesion segmentations from a subset of five cases with mild to moderate lesion burden were evaluated at the lesion-level to determine which lesions were captured in the ULF reference mask. Spatial clustering of lesions does not provide a 1:1 mapping between WMLs on ULF and 3T images and cannot differentiate confluent lesions that are spatially connected. To address this, A.T. selected individual lesions on the 3T FLAIR images, deliberately avoiding large confluent lesions. For each individual lesion, 3T lesion volumes were extracted, and A.T. visually categorised lesion locations as deep white matter (DWM), periventricular (PV), juxtacortical (JC), brainstem, or cerebellum. Each lesion was classified as *ULF detected* if included in the ULF Aided mask or *ULF missed* if it was not included in the ULF Aided mask.

4.1.2.5 Evaluation of Automated Ultra-Low Field Segmentation Approaches

Lesion volumes from each automatic method were compared with those from the ULF Aided masks, ULF Blinded masks, and 3T reference masks. Volumes were calculated in the original output space of each segmentation mask. The performance of the automatic methods was assessed by calculating percent error, volume correlation, and the intraclass correlation coefficient (ICC) relative to reference masks.

For all automatic 64mT WML segmentation algorithms, overlap metrics were evaluated against the ULF Aided masks. Overlap metrics were extracted using FSL's *bianca_overlap_tools*, following the definitions for all overlap metrics (DSC, voxel- and cluster-level false positive ratio, voxel- and cluster-level false negative ratio, detection error rate, outline error rate) as described in the original paper (Griffanti et al., 2016). Only participants with ULF Aided segmentation volume > 0 were included. In the subset of participants with ULF Blinded masks, the accuracy of the automated segmentations relative to ULF Blinded masks was assessed. The ULF Aided masks were still used as a reference in this comparison, and the same overlap metrics were applied.

As SynthSeg's output segmentations are returned at a resolution of 1 mm isotropic, to calculate overlap metrics, SynthSeg WML masks were down-sampled to the original ULF voxel size. Since down-sampling could disadvantage the SynthSeg output, the analysis was re-performed in the original output space of SynthSeg (1×1×1 mm). Aided ULF masks and all other segmentation outputs were up-sampled to 1×1×1 mm for this analysis. To minimise resampling errors and preserve the original lesion volume, all resampled (up-sampled or down-sampled) WML masks were linearly interpolated and thresholded at a value that minimised the difference in volume relative to the original lesion mask volume, selected using iterative thresholds in steps of 0.05.

In a subset of 24 participants, an experienced neurologist (S.V.O.) assessed the quality of WML segmentations produced by MIMoSA, SynthSeg, nnU-Net FLAIR, nnU-Net T1w, PLAn 2 FLAIR, and PLAn 2 FLAIR/T1w. The 24 participants were randomly selected to ensure equal representation across each lesion burden group (mild, moderate, or severe). For each participant, anonymised labels were randomly assigned to each segmentation method, and a 3T FLAIR scan was provided as a reference to assist in evaluating the segmentation quality. The ratings were based on a scale from 1 to 5, with 5

representing the highest quality segmentation (Table 4.3). The mean score and 95% confidence intervals were reported for each method.

Table 4.3: Qualitative Rating Scale Definitions for Automatic Segmentation Methods at 64mT

Score	Rating	Definition
1	Very Poor	0–25% of WML visible on ULF MRI segmented; 100–75% not segmented or other structures segmented as WML
2	Poor	25–50% of WML visible on ULF MRI segmented; 75–50% not segmented or other structures segmented as WML
3	OK	50–75% of WML visible on ULF MRI segmented; 50–25% not segmented or other structures segmented as WML
4	Good	75–100% of WML visible on ULF MRI segmented; 25–0% not segmented or other structures segmented as WML. In the absence of ULF visible WML, minimal false positive identification
5	Excellent	Almost all lesions visible on ULF MRI segmented, well defined and matching the shape of the lesions; minimal false positive areas

Abbreviations: WML = white matter lesion; ULF = ultra-low field

4.1.2.6 Optimal Thresholding of Ultra-Low Field White Matter Lesion Probability Maps

As outlined in Table 4.2, the MIMoSA, PLAn, and nnU-Net models generate probability maps indicating the likelihood of each voxel being a lesion. The binarised WML masks used for the main analysis were based on the default thresholds for each model: 0.2 for MIMoSA, as specified in the original publication, and 0.5 for all the PLAn and nnU-Net models. To investigate how the probability map threshold affects segmentation performance, the overlap analyses for these models were re-run across a range of thresholds from 0 to 1, with increments of 0.05.

4.1.2.7 Analysis of Results across ULF Acquisition Parameters

The ULF dataset consisted of scans acquired with heterogeneous acquisition parameters, as detailed in the appendix (Appendix Table A.4.1). In an exploratory analysis, the effects of ULF software version, ULF FLAIR in-plane voxel size, and ULF FLAIR reconstruction method on ULF automatic segmentation performance were considered. For each category, the validation cohort was split into subgroups, and I examined whether the general trends of DSC for automatic segmentation methods relative to ULF Aided reference masks remained consistent across the subgroups.

Given that factors such as total lesion burden may influence segmentation performance, differences in total white matter volume between subgroups, measured using reference 3T masks or ULF-aided masks, were also assessed. The 3T scans, which had homogeneous acquisition parameters, provided a basis for comparing total lesion volume across subgroups, irrespective of how ULF acquisition parameters might have affected the total white matter volume.

Due to the retrospective nature of the dataset, the number of participants in each subgroup within acquisition categories was unequal, and there were unbalanced participant characteristics across subgroups. Due to the lack of a controlled dataset, where the same participants were scanned with different parameter settings, the direct effects of image acquisition settings on segmentation performance could not be assessed. As a result, I focused on evaluating visual trends rather than performing formal statistical analyses.

4.1.2.8 Statistical Analysis

Statistical analyses were conducted in GraphPad Prism 10.1.0. Data were checked for normality using the D'Agostino & Pearson test. Differences between categorical and continuous variables in the training and validation cohorts were assessed with two-tailed Fisher's Exact and Mann-Whitney U tests, respectively.

After assessing normality assumptions, volumetric correspondence between 3T and ULF reference masks, or between segmentation methods and reference masks, was quantified using Spearman's correlation, percentage error (computed as the absolute difference in volume between two segmentations divided by the mean of the two volumes), and ICC employing a two-way mixed model with absolute agreement. Additionally, when comparing 3T and 64mT reference masks, a linear model was fit to 3T lesion count and log-transformed 3T WML volume to evaluate the effect of these factors on DSC between ULF Aided and 3T reference masks.

Segmentation performance outcome measures (volume, overlap metrics, lesion count, and qualitative ratings) across segmentation methods were evaluated using a t-test or the Friedman test, followed by a post-hoc Dunn's multiple comparisons test. A corrected p-value of 0.05 or lower was considered statistically significant. Unless otherwise noted, all data reported are in the form of mean ± standard error.

4.1.3 Results

4.1.3.1 Evaluation: Reference Segmentations

Total WML volume and lesion count were calculated in native image space. Total ULF Aided WML volume ($8049 \pm 1117\,\text{mm}^3$) was higher than 3T WML volume ($6308 \pm 970\,\text{mm}^3$), with an average $16 \pm 32\%$ higher ULF WML volume than 3T ($Z = 2464$, $p < 0.0001$, Figure 4.1A). Total lesion count was significantly higher at 3T (24.6 ± 1.9) compared to ULF Aided masks (9.1 ± 0.68), with an average of 15.5 ± 13.8 more WML clusters detected on 3T than ULF Aided ($Z = 3481$, $p < 0.001$, Figure 4.1B). The larger voxel size in ULF segmentation masks likely drives the increase in volume seen at ULF, even if lesions are undetected at ULF. Both volume and lesion count were strongly correlated between ULF Aided and 3T ($r_s(82) = 0.93$, $p < 0.0001$, and $r_s(82) = 0.81$, $p < 0.0001$) (Table 4.4).

The mean DSC of ULF Aided masks compared to 3T reference masks was 0.30 ± 0.02, ranging from 0 up to 0.65. DSC was higher in cases with larger total lesion volumes: 3T WML log volume showed a strong positive correlation with DSC ($r_s(82) = 0.86$, $p < 0.0001$, Figure 4.1C). The total 3T lesion count had a weaker, but still positive, correlation with DSC ($r_s(82) = 0.55$, $p < 0.0001$).

In a subset of 31 participants, ULF Blinded WML masks were drawn on ULF FLAIR images without a 3T reference scan. This subset had a slightly lower average 3T total WML volume ($5976 \pm 1676\,\text{mm}^3$) compared to the entire cohort ($6308 \pm 970.3\,\text{mm}^3$), but this difference was not significant ($U = 1340$, $p = 0.45$). ULF Blinded total WML volume ($4420 \pm 1269\,\text{mm}^3$) was significantly lower than both ULF Aided ($7742 \pm 1912\,\text{mm}^3$, $Z = 4.9$, $p < 0.0001$) and 3T volume ($5976 \pm 1676\,\text{mm}^3$, $Z = 2.8$, $p < 0.015$), but there was no significant difference between ULF Aided and 3T volume ($Z = 2.1$, $p < 0.11$) (Figure 4.1D). Total lesion count was higher for 3T (17 ± 3.0) than both ULF Aided (7 ± 1, $Z = 6.9$, $p < 0.0001$) and ULF Blinded (11 ± 1.6, $Z = 4.5$, $p < 0.0001$). However, there was no significant difference between ULF Aided and ULF Blinded ($Z = 2.4$, $p = 0.051$) (Figure 4.1E). ULF Aided had a slightly higher but not significantly different DSC score (0.30 ± 0.03) compared to ULF Blinded (0.20 ± 0.03, $df = 32$, $t = 1.7$, $p = 0.10$) (Figure 4.1F).

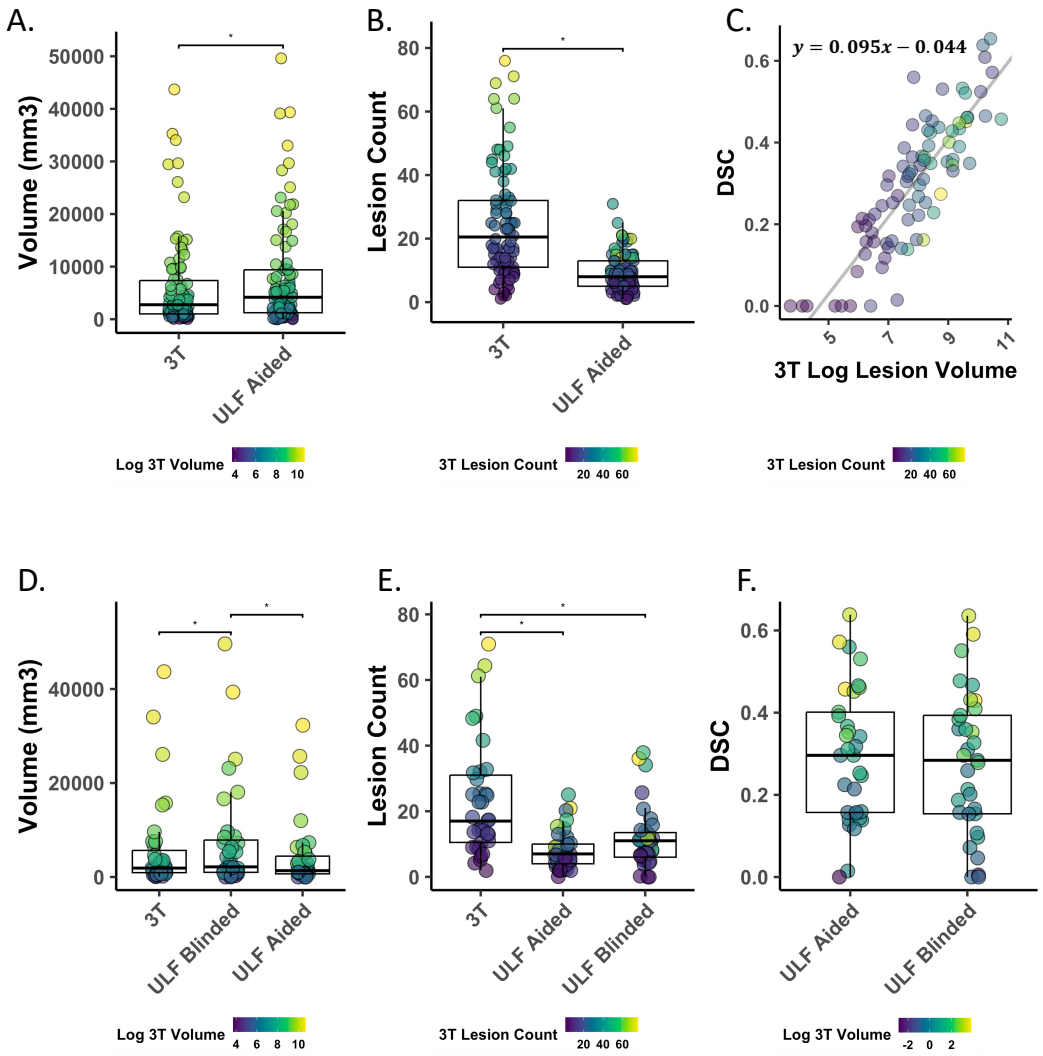


Figure 4.1: Comparison of 3T and ULF manual lesion segmentations. Subplots A–C show data from the validation cohort (n=84), and subplots D–F show a subset with ULF Blinded segmentation masks (n=35). A) ULF Aided WML volume compared with 3T reference volume. B) ULF Aided WML lesion count compared with 3T reference volume. C) Scatter plot of DSC of up-sampled ULF Aided relative to log-transformed 3T reference volumes. D) ULF Aided and ULF Blinded WML volumes compared with 3T reference. E) ULF Aided and ULF Blinded lesion counts compared with 3T reference. F) DSC of ULF Aided and Blinded segmentations relative to 3T reference. Box plots show median, 25th, and 75th percentiles, with individual participant values.

Table 4.4: Overview of all volume comparison between automatic segmentation masks with reference masks at HF and ULF.

Algorithm	**Ref. Seg.**	**Val. Dataset**	**Spearman Corr.**		**APE**		**ICC**			
			rho	p-value	Avg.	SD	ICC	UB	LB	p-value
ULF Aided	3T	All	0.97	< 0.001	50	47	0.93	0.85	0.96	< 0.001
ULF Aided	3T	Blinded subset	0.94	< 0.001	51	48	0.93	0.86	0.97	< 0.001
ULF Blinded	3T	Blinded subset	0.94	< 0.001	60	57	0.93	0.83	0.97	< 0.001
MIMoSA	ULF Aided	All	0.41	< 0.001	115	60	0.31	-0.08	0.60	0.07
SynthSeg	ULF Aided	All	0.58	< 0.001	81	60	0.70	0.55	0.78	< 0.001
nnUNet FL	ULF Aided	All	0.80	< 0.001	93	65	0.69	0.40	0.83	< 0.001
nnU-Net FL/T1w	ULF Aided	All	0.81	< 0.001	93	67	0.81	0.54	0.90	< 0.001
PLAn 1 FL	ULF Aided	All	0.90	< 0.001	68	64	0.93	0.86	0.97	< 0.001
PLAn 1 FL/T1w	ULF Aided	All	0.90	< 0.001	67	62	0.92	0.85	0.96	< 0.001
PLAn 2 FL	ULF Aided	All	0.90	< 0.001	62	64	0.92	0.85	0.95	< 0.001
PLAn 2 FL/T1w	ULF Aided	All	0.90	< 0.001	65	63	0.93	0.87	0.96	< 0.001
MIMoSA	3T	All	0.41	< 0.001	128	55	0.27	-0.20	0.57	0.11
SynthSeg	3T	All	0.60	< 0.001	91	59	0.71	0.45	0.85	< 0.001
nnUNet FL	3T	All	0.79	< 0.001	81	65	0.77	0.63	0.86	< 0.001
nnUNet FL/T1w	3T	All	0.82	< 0.001	82	65	0.89	0.81	0.94	< 0.001
PLAn 1 FL	3T	All	0.90	< 0.001	60	55	0.96	0.936	0.97	< 0.001
PLAn 1 FL/T1w	3T	All	0.89	< 0.001	58	58	0.96	0.94	0.97	< 0.001
PLAn 2 FL	3T	All	0.89	< 0.001	60	54	0.95	0.92	0.96	< 0.001
PLAn 2 FL/T1w	3T	All	0.89	< 0.001	63	56	0.94	0.91	0.96	< 0.001
ULF Blinded	ULF Aided	Blinded subset	0.87	< 0.001	61	48	0.82	0.58	0.92	< 0.001
MIMoSA	ULF Aided	Blinded subset	0.37	0.031	127	55	0.41	-0.10	0.73	0.08
SynthSeg	ULF Aided	Blinded subset	0.53	0.001	91	62	0.56	0.35	0.78	< 0.001
nnU-Net FL	ULF Aided	Blinded subset	0.63	< 0.001	101	72	0.75	0.48	0.88	< 0.001
nnU-Net FL/T1w	ULF Aided	Blinded subset	0.85	< 0.001	105	72	0.82	0.60	0.92	< 0.001
PLAn 1 FL	ULF Aided	Blinded subset	0.84	< 0.001	85	67	0.92	0.83	0.96	< 0.001
PLAn 1 FL/T1w	ULF Aided	Blinded subset	0.87	< 0.001	77	69	0.90	0.80	0.95	< 0.001
PLAn 2 FL	ULF Aided	Blinded subset	0.87	< 0.001	73	68	0.91	0.820	0.95	< 0.001
PLAn 2 FL/T1w	ULF Aided	Blinded subset	0.89	< 0.001	73	72	0.91	0.83	0.96	< 0.001

Abbreviations: Ref. Seg. = reference segmentation; Val Dataset = validation dataset; APE = average percent error; Intraclass Correlation Coefficient; Avg. = Average; SD = standard deviation; UB = upper boundary; LB = lower boundary; ULF = ultra-low field
Number of participants in validation datasets: All ($n = 84$) and Blinded subset ($n = 35$)

In a subset of 126 non-confluent lesions across five participants, individual 3T WML volumes and lesion locations were recorded, and it was visually determined whether each lesion was detected in the ULF Aided mask. This analysis corroborated quantitative findings, revealing that, compared to 3T segmentation, many lesions were missed on ULF Aided masks, while detected lesions had overestimated volumes, as shown in the example in Figure 4.2A. Of the 126 lesions, 52 (41%) were detected on ULF. Lesions detected on ULF had significantly larger 3T volumes (235.2 ± 35.53) than those missed (31.9 ± 4.2, $U = 218$, $p < 0.0001$, two-tailed) (Figure 4.2B). Notably, 79.3% of lesions

larger than 50 mm^3 were detected on ULF, whereas only 7.3% of the 68 lesions smaller than 50 mm^3 were detected.

The location of WMLs also influenced ULF detectability (Figure 4.2C–D). For instance, infratentorial lesions were missed on ULF, which may have been affected by poorer image quality due to the geometry of the scanner, leading to severe signal loss away from its centre. PV lesions were the most discernible on ULF, with 63% of the 64 PV lesions detected, while only around 20% of DWM and JC lesions were detectable on ULF.

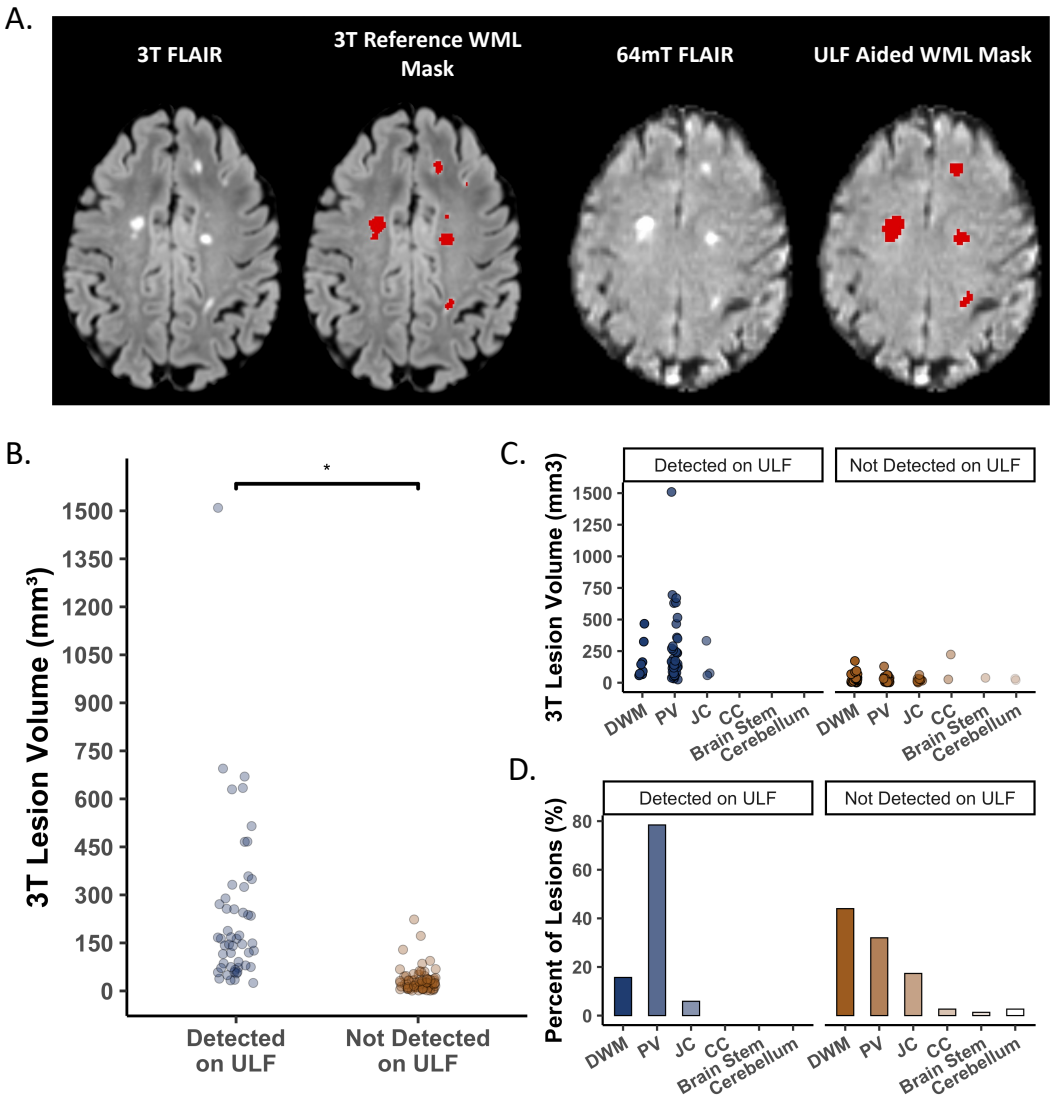


Figure 4.2: Visual detection of individual WMLs on ULF. In 5 participants, 102 non-confluent lesions were selected on 3T MRI and visually marked as detected (blue) or not detected (orange) on ULF Aided segmentations. A) Example paired 3T-64mT FLAIR scans with reference WML lesions. B) Lesions successfully detected on ULF had higher volumes than those not detected (U=218, p¡0.0001, two tailed). C) Volume of individual lesions detected/missed on ULF across different brain locations. D) Percentage of individual lesions at each location for lesions detected/missed by ULF. *WML locations: DWM (Deep White Matter), PV (Periventricular), JC (Juxtacortical), CC (Corpus Callosum).

4.1.3.2 Quantitative Evaluation: Automated Ultra-Low Field Segmentations

Total WML lesion volumes obtained with automated approaches were compared to 3T and ULF reference segmentations. As shown in Figure 4.3A, total 3T WML volume ($6308 \pm 970\,\mathrm{mm}^3$) was significantly lower than segmentation outputs from MIMoSA ($21259 \pm 1356\,\mathrm{mm}^3$, $z = 9.1$, $p < 0.0001$) and SynthSeg ($9253 \pm 567\,\mathrm{mm}^3$, $z = 5.0$, $p < 0.0001$). In contrast, both nnU-Net models produced significantly lower total WML volumes (nnU-Net FLAIR: $4309 \pm 767\,\mathrm{mm}^3$ and nnU-Net FLAIR/T1w: $4806 \pm 812.6\,\mathrm{mm}^3$, $z = 6.0$, $p < 0.0001$ for both) compared to 3T.

The 3T volume showed weak but significant correlation with MIMoSA ($r_s(82) = 0.4$, $p < 0.001$) (Table 4.4). All PLAn models exhibited a strong correlation with 3T volumes ($r_s(82) > 0.94$), similar to the correlation strength between 3T and ULF Aided volumes ($r = 0.93$, $p < 0.0001$).

When compared to ULF Aided volumes ($8049 \pm 1118\,\mathrm{mm}^3$), only MIMoSA had a significantly higher volume ($z = 6.2$, $p < 0.001$), while SynthSeg ($9253 \pm 567\,\mathrm{mm}^3$, $z = 1.9$, $p = 0.49$), PLAn 2 FLAIR ($6502 \pm 932.4\,\mathrm{mm}^3$, $z = 2.3$, $p = 0.19$), and PLAn 2 FLAIR/T1w ($6655 \pm 990\,\mathrm{mm}^3$, $z = 2.6$, $p = 0.07$) did not show significant differences. All automatic models demonstrated significant correlations between volume and ULF Aided volume. MIMoSA showed the weakest correlation ($r_s(82) = 0.4$, $p < 0.001$), and PLAn models showed the strongest volume correlations (Table 4.4).

In the blinded validation cohort subset ($n = 31$), the relationships between ULF Aided volumes and automated segmentation methods (Figure 4.3C) mirrored the findings from the entire validation cohort (Figure 4.3B). PLAn models were between 1268–$1614\,\mathrm{mm}^3$ higher than ULF Blinded volumes ($4420 \pm 1269\,\mathrm{mm}^3$) but showed no significant difference in total WML volumes. WML volumes from MIMoSA and SynthSeg segmentation outputs were significantly higher than ULF Blinded volumes ($z = 8.0$, $p < 0.0001$ and $z = 5.5$, $p < 0.0001$).

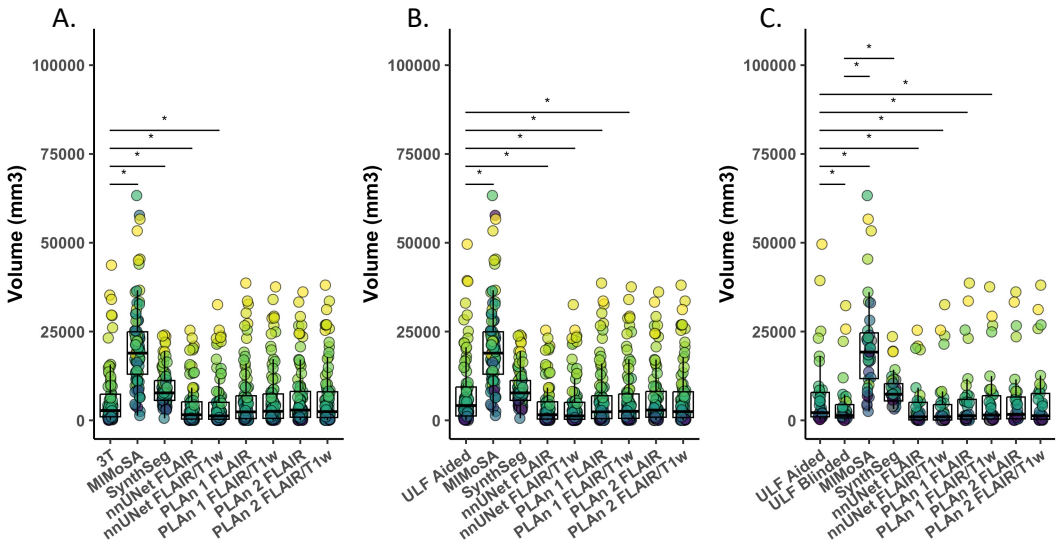


Figure 4.3: Comparison of total WML volume across automatic segmentation methods to A) 3T manual segmentation mask, B) ULF Aided segmentation mask, and C) ULF Aided and ULF Blinded segmentation mask in validation cohort subset (n=31). Box plots show median, 25th and 75th percentiles, with individual participant values. Significant differences (p¡0.05) from Friedman test with Dunn's multiple comparisons are marked (*).

In the validation cohort ($n = 84$), WML overlap metrics were calculated for all automated segmentation methods with the ULF Aided manual masks as the reference. Four participants were excluded from this analysis who had a total WML volume of 0 in the ULF Aided reference masks.

All methods demonstrated higher DSC with higher total lesion volumes (Figure 4.4A). The PLAn models exhibited the highest mean DSC scores, ranging from 0.48 to 0.49, with no significant differences between the models, but significantly higher than all other methods. MIMoSA had the lowest DSC (0.23 ± 0.02), significantly lower than all other methods. SynthSeg had the second-lowest DSC (0.30 ± 0.03), with around 0.10 lower mean DSC than both nnU-Net models, but not significantly different from nnU-Net FLAIR (0.41 ± 0.03, $z = 2.2$, $p = 0.72$).

The nnU-Net models were the most conservative, yielding a low false positive ratio (FPR) and a high false negative ratio (FNR). Indeed, the nnU-Net models had the lowest mean FPR (nnU-Net FLAIR = 0.2 ± 0.03, nnU-Net FLAIR/T1w = 0.2 ± 0.03), which were significantly lower than all other models, except for one comparison showing no difference between nnU-Net FLAIR and PLAn 1 T1w/FLAIR (FPR = 0.3 ± 0.03, $z = 3.1$, $p = 0.05$). The nnU-Net models also had the highest FNR (nnU-Net FLAIR: 0.66 ± 0.3, nnU-Net FLAIR/T1w: 0.66 ± 0.03) and the highest FNR clusters (nnU-Net FLAIR

$= 0.53 \pm 0.3$, nnU-Net FLAIR/T1w $= 0.55 \pm 0.03$) compared to all other methods (Figure 4.4A). Conversely, MIMoSA showed the highest FPR (0.81 ± 0.02) and FPR clusters (0.9 ± 0.01), followed by SynthSeg (FPR $= 0.70 \pm 0.03$, FPR Cluster $= 0.60 \pm 0.2$). PLAn 2 FLAIR had lower FNR than SynthSeg ($z = 3.3$, $p = 0.03$), while other PLAn models showed similar FNR to SynthSeg and MIMoSA. MIMoSA's tendency for over-segmentation likely contributes to its better cluster detection ability, as its FNR cluster was lower than all PLAn models, despite no significant difference in voxel-level FNR.

PLAn 1 FLAIR, PLAn 2 FLAIR, and PLAn 2 FLAIR/T1w were the best models at detecting WML, with mean DERs of 0.47. These models also showed statistically similar mean OER, ranging from 0.54 to 0.57, much lower than SynthSeg's mean OER of 0.90 ± 0.03. MIMoSA had the highest mean DER (1.23 ± 0.065), indicating that the total detection errors exceeded the mean total WML volume (of MIMoSA and ULF Aided). However, MIMoSA's mean OER was significantly lower than all other methods (0.30 ± 0.03), suggesting that the few lesions it managed to detect correctly were well outlined compared to ULF Aided.

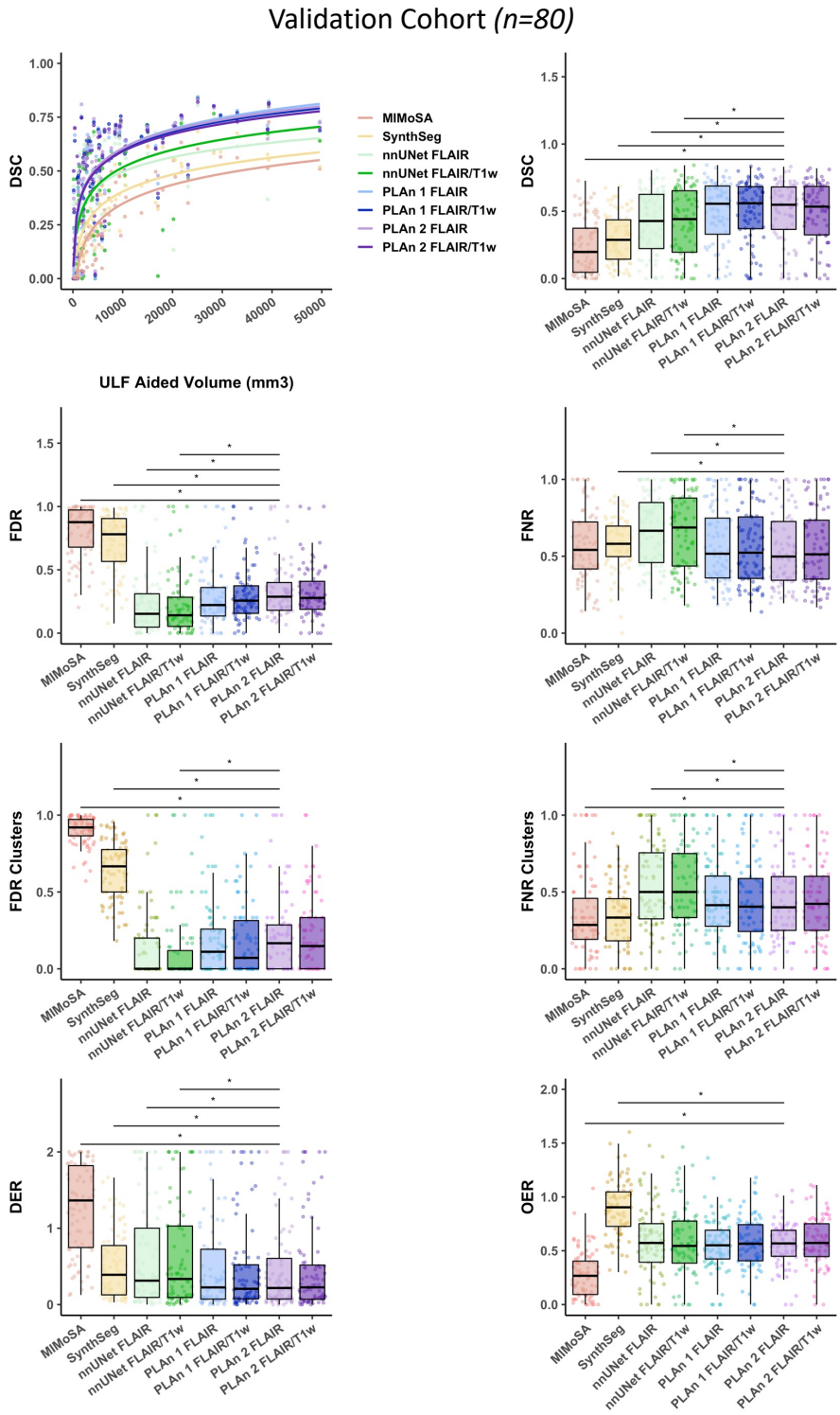


Figure 4.4: Evaluation of automatic WML segmentation performance in original ULF resolution. Overlap metrics of automatic segmentation algorithms using ULF Aided Manual masks as a reference. Scatter plots of DSC vs ULF Aided volume are fitted with logarithmic non-linear regression for each method. Overlap metrics in the entire validation cohort (n=80). Significant differences are only shown between PLAn 2 FLAIR and other methods. For all post-hoc pairwise comparisons see Appendix D.1.1.

In Figure 4.5, the overlap analysis was repeated on the subset of cases with Blinded ULF masks, only including participants with ULF Aided volume greater than zero ($n = 31$). The overall trend in the overlap metrics between the automated segmentation methods remained consistent for both the validation cohort (Figure 4.4) and the validation subset (Figure 4.5). However, some statistical significance changed due to the reduced statistical power in the subset analysis (Appendix D.1.1).

All PLAn models performed similarly to the Blinded ULF manual segmentation, with no significant differences between any PLAn model and Blinded ULF for any overlap metric. SynthSeg generally performed well relative to Blinded ULF, exhibiting only a significantly higher FPR, likely due to over-segmentation near the ventricles. In contrast, MIMoSA showed poor segmentation performance compared to Blinded ULF across all overlap metrics.

Figures 4.4 and 4.5 show overlap metrics calculated in the native ULF resolution, with SynthSeg masks down-sampled from their original $1 \times 1 \times 1$ mm output. To account for potential bias of resampling SynthSeg, the analysis was repeated at $1 \times 1 \times 1$ mm resolution (Appendix D.1.1). The trends were largely consistent across resolutions, though some pairwise statistical differences varied (Appendix D.1.1).

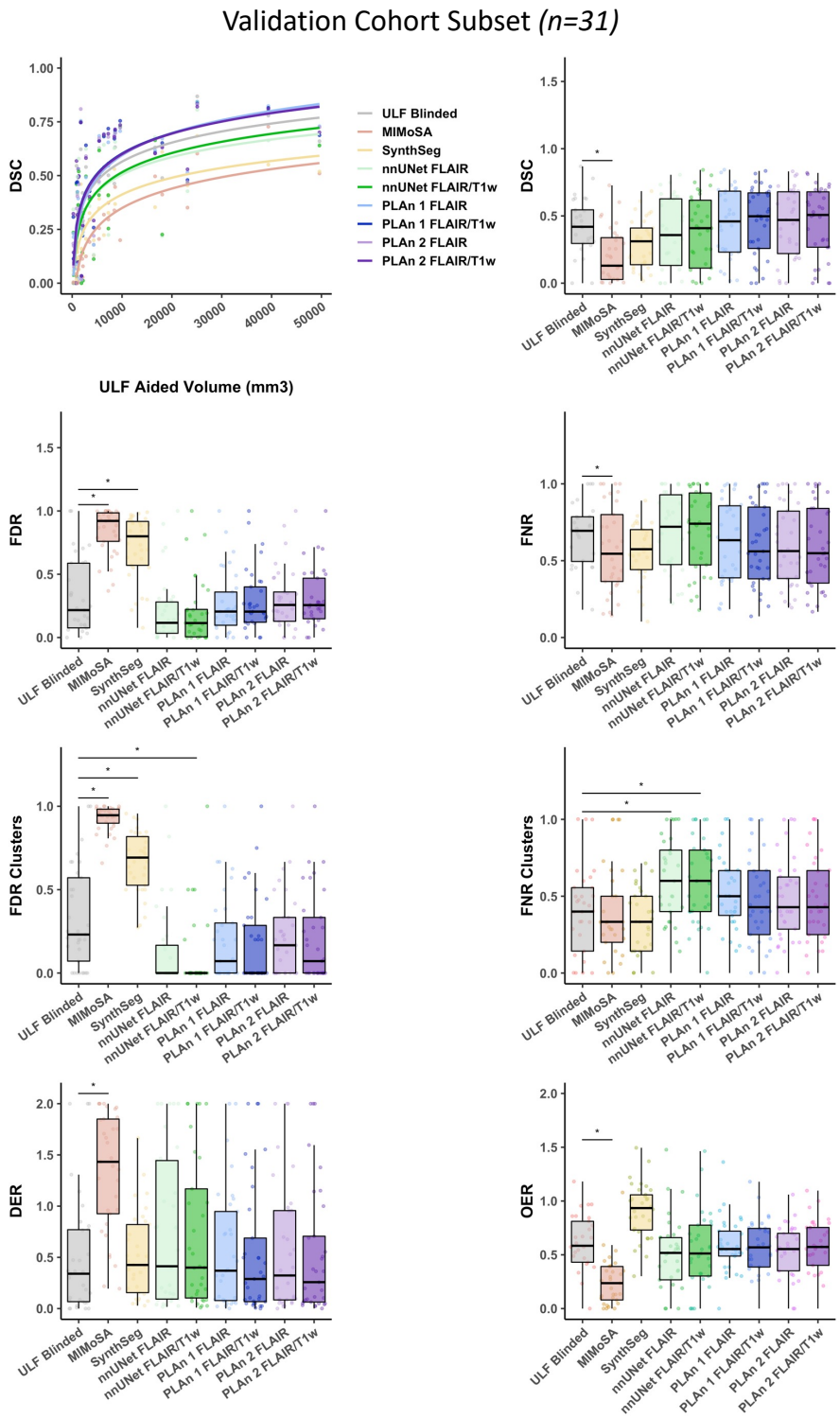


Figure 4.5: Evaluation of automatic WML segmentation performance in original ULF resolution. Overlap metrics of automatic segmentation algorithms using ULF Aided Manual masks as a reference. Scatter plots of DSC vs ULF Aided volume are fitted with logarithmic non-linear regression for each method. Overlap metrics in the subset of the validation cohort with ULF Blinded masks ($n = 31$). Significant differences are only shown between ULF Blinded and other methods.

4.1.3.3 Qualitative Evaluation: Automated Ultra-Low Field Segmentation

The quality of WML segmentations was assessed by an experienced neurologist using a five-point rating scale, where higher scores indicated better segmentation quality (Table 4.3). Only six of the methods were included in the qualitative ratings: MIMoSA, SynthSeg, nnU-Net FLAIR, nnU-Net FLAIR/T1w, PLAn 2 FLAIR, and PLAn 2 FLAIR/T1w. Since PLAn models performed similarly in the overlap metrics, PLAn 1 FLAIR and PLAn 1 FLAIR/T1w were not included in the qualitative ratings. Table 4.5 summarises the qualitative ratings.

PLAn outputs were most similar to the ULF Aided masks (Figure 4.6). A Friedman test revealed significant differences among the methods ($p < 0.0001$), with MIMoSA receiving the lowest mean score (1.92 ± 0.15) and PLAn 2 FLAIR/T1w achieving the highest (4.08 ± 0.15). MIMoSA performed significantly worse than nnU-Net FLAIR (3.17 ± 0.20, $Z = 3.8$, $p = 0.0023$), nnU-Net FLAIR/T1w (3.46 ± 0.15, $Z = 5.1$, $p < 0.0001$), PLAn 2 FLAIR (3.96 ± 0.16, $Z = 6.3$, $p < 0.0001$), and PLAn 2 FLAIR/T1w ($Z = 6.8$, $p < 0.0001$). In multiple cases, SynthSeg incorrectly marked hypointense regions of the FLAIR as lesions (Figure 4.6, case 2), contributing to its low ratings (2.79 ± 0.15). The SynthSeg rating was not significantly different from MIMoSA ($Z = 2.6$, $p = 0.15$), but significantly lower than PLAn 2 FLAIR ($Z = 3.742$, $p = 0.0027$) and PLAn 2 FLAIR/T1w ($Z = 4.205$, $p = 0.0004$). Among the nnU-Net-based methods, nnU-Net FLAIR/T1w performed significantly better than nnU-Net FLAIR ($Z = 3.0$, $p = 0.0393$). However, no significant differences were observed between PLAn 2 FLAIR and PLAn 2 FLAIR/T1w ($Z = 0.5$, $p > 0.99$).

Table 4.5: Qualitative Ratings of Automatic WML Segmentations

Algorithm	**WML Rating (mean ± CI)**
MIMoSA	1.9 [1.6, 2.2]
SynthSeg	2.8 [2.5, 3.1]
nnU-Net FLAIR	3.2 [2.8, 3.6]
nnU-Net FLAIR/T1w	3.5 [3.2, 3.8]
PLAn 2 FLAIR	4.0 [3.6, 4.3]
PLAn 2 FLAIR/T1w	4.1 [3.8, 4.4]

Abbreviations: WML = white matter lesion

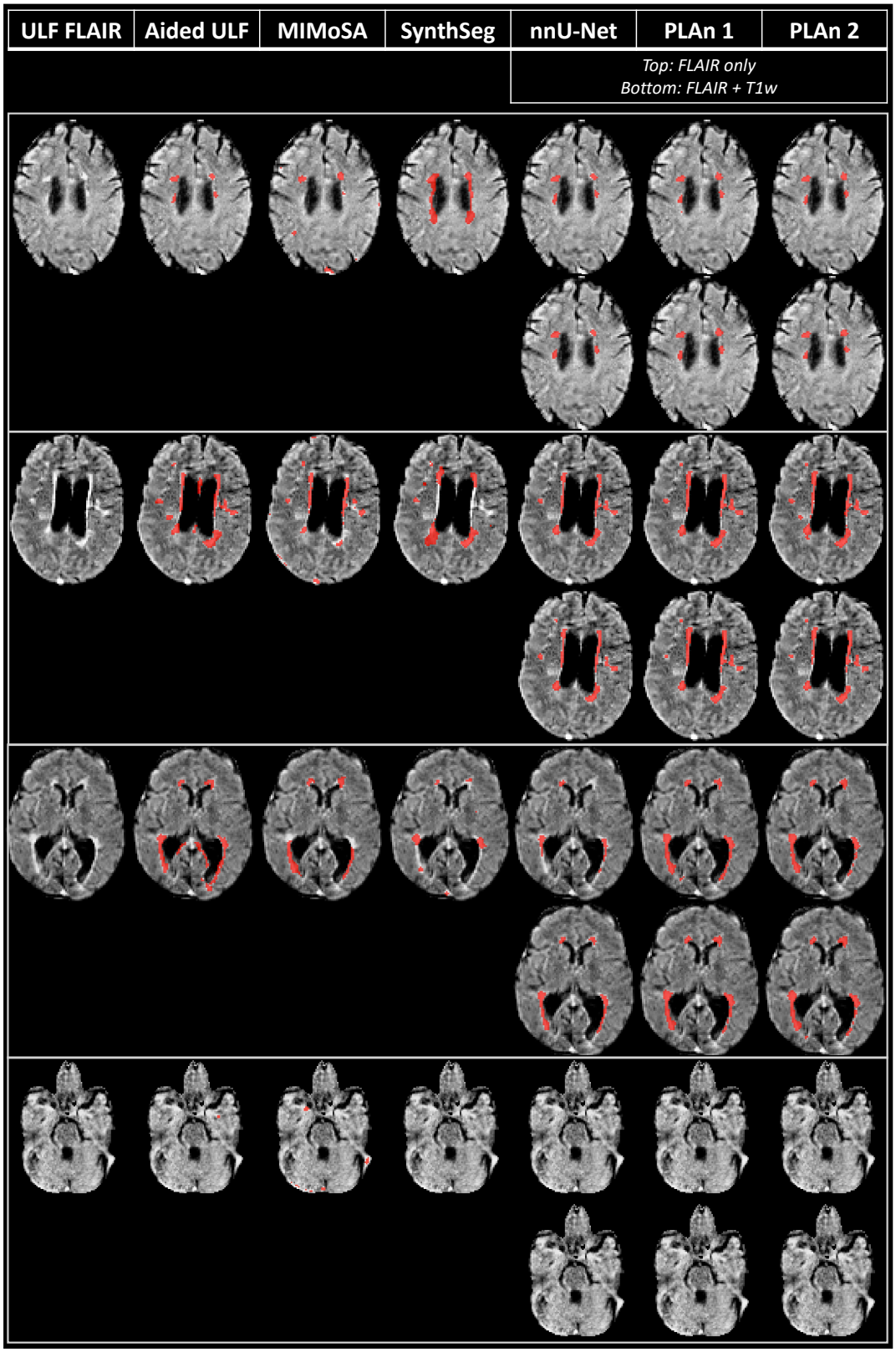


Figure 4.6: Example WML masks for ULF Aided Reference and automatic methods overlaid on ULF FLAIR. Different cases are contained in white boxes, with nnU-Net and PLAn models trained on FLAIR only in the top row, and FLAIR+T1w on the bottom row.

4.1.3.4 Optimal Thresholding of Ultra-Low Field White Matter Lesion Probability Maps

MIMoSA's default output is a probability map, requiring thresholding to generate a binary mask (Figure 4.7A). In the original MIMoSA publication, a probability threshold of 0.2 was determined to be optimal for segmentation performance and has been used in all analyses up to this point (Valcarcel et al., 2018). However, this threshold was based on 3T data and may not be optimal when applied to ULF data. Additionally, the nnU-Net package has an option to save the SoftMax probability outputs, which can also be applied to save probability maps for all PLAn and nnU-Net models. The impact of different probability thresholds on segmentation performance was then systematically evaluated by calculating DSC at incremental thresholds of 0.05 for MIMoSA, nnU-Net, and PLAn models, using the ULF Aided mask as the reference segmentation.

For MIMoSA, the highest mean DSC was observed at a threshold of 0.25, yielding a DSC of 0.24 ± 0.02. This was not significantly different from the default DSC of 0.23 ± 0.02 ($z = 1.03$, $p > 0.99$, Figure 4.7). Similarly, for PLAn and nnU-Net models, the default segmentation at 0.5 was not significantly lower than the optimal threshold for this cohort (Figures 4.7C–H). Both nnU-Net models had optimal thresholds at 0.1, resulting in 0.07 higher DSC scores. DSC for PLAn models were quite stable across all thresholds.

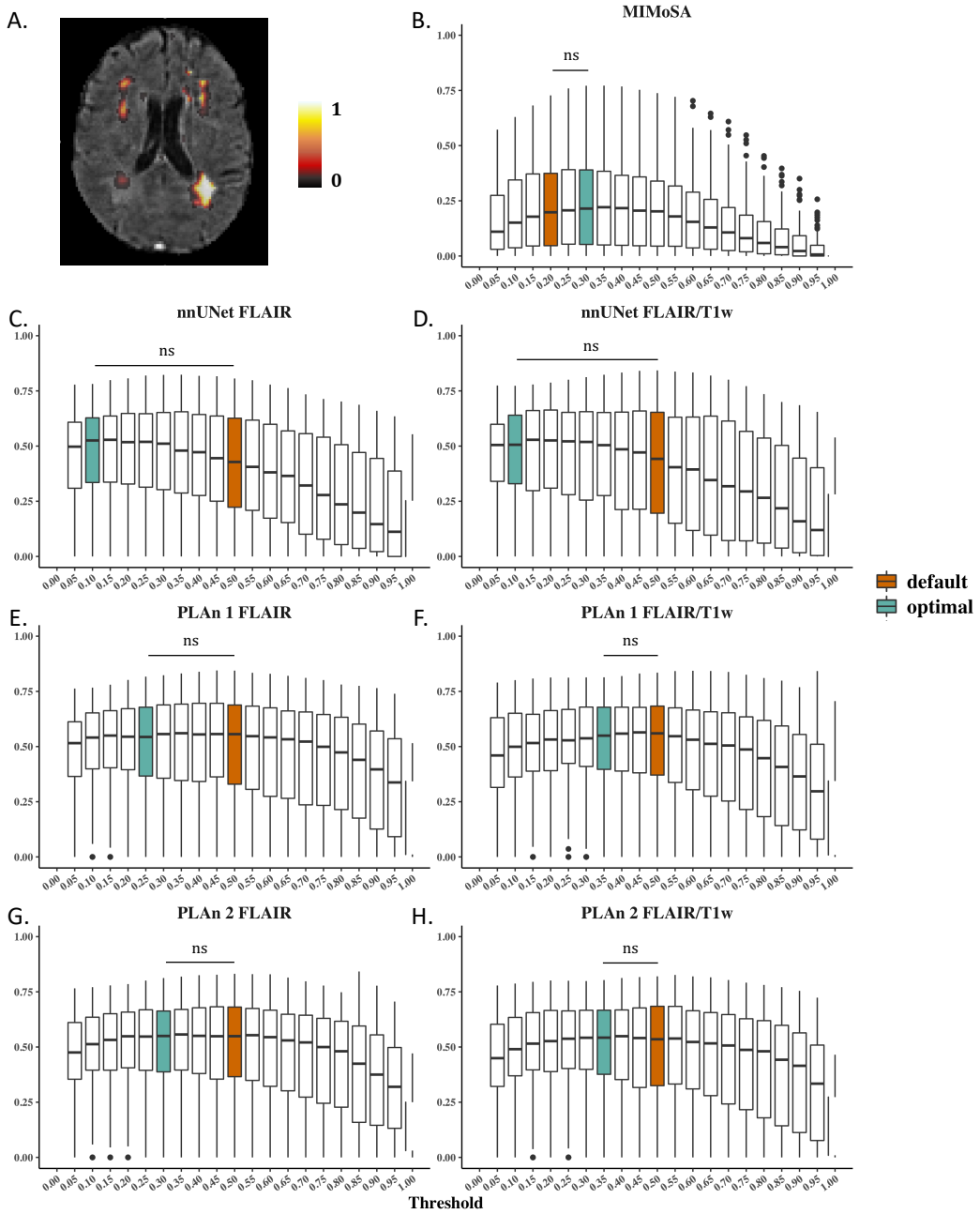


Figure 4.7: Effects of probability mask thresholding on DSC for MIMoSA, nnU-Net, and PLAn models. A) An example WML probability mask, showing probabilities from 0-1 of lesions at candidate voxel. B-H) Panels for each method that outputs a probability mask. The DSC values were calculated at incremental probability thresholds (0.05 intervals) using the ULF Aided mask as the reference. The orange highlight indicates the default threshold used in the previous analysis, while the teal highlight shows the threshold at which the mean DSC is maximized. A Friedman test with multiple comparisons (Dunn's test) found no significant differences in DSC between the optimal and default thresholds for any of the methods.

When comparing overlap metrics across automatic segmentation algorithms using these optimal (although biased) thresholds, MIMoSA still had significantly worse overlap scores compared to PLAn models (Figure 4.8). However, nnU-Net models performed more similarly to PLAn (Figure 4.8).

These findings suggest that while default segmentations are generally optimal for all models, nnU-Net's under segmentation might be partially compensated for by thresholding at a lower probability value.

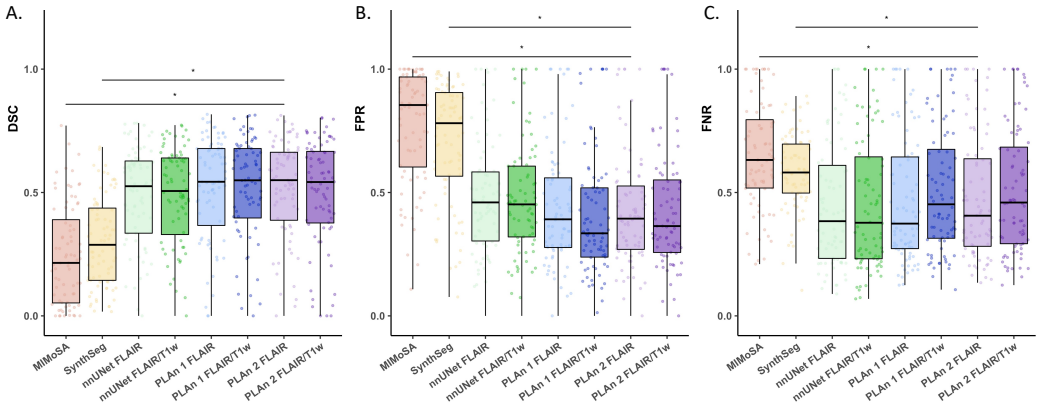


Figure 4.8: Comparison of overlap metrics across different models, using optimized probability thresholds for MIMoSA, nnU-Net and PLAn models. A) DSC, B) FPR, C) FNR for optimized probability mask thresholds, calculated at the threshold that yields the highest mean DSC for each model. Statistically significant pairwise comparisons are shown only between the PLAn 2 FLAIR model and the other methods

4.1.3.5 Analysis of Results across ULF Acquisition Parameters

The results of acquisition heterogeneity on segmentation performance are presented in Appendix D.1.2. Segmentation performance varied by model and scan parameters, with PLAn consistently achieving high accuracy across all sub-cohorts, nnU-Net showing sensitivity to software version and resolution, and SynthSeg demonstrating stable performance across settings but lower accuracy than PLAn (Appendix D.1.2).

4.1.4 Discussion

In this section, I present a comprehensive analysis of the performance of automatic WML segmentation methods at ULF. SynthSeg, nnU-Net models, and PLAn models, which are automatic segmentation methods designed for ULF data, performed comparably to ULF Blinded masks. The reference ULF Blinded masks were manually annotated without reference to corresponding 3T scans, highlighting the clinical utility of the automatic methods, especially when a 3T reference is not available. Additionally, PLAn models adapted for our ULF data have superior WML segmentation performance when evaluated quantitatively and qualitatively compared to other segmentation methods. While nnU-Net models trained only on ULF data could achieve comparable

overlap metrics to PLAn when their probability masks were thresholded at a lower value, PLAn models were more stable across probability thresholds and ULF acquisition parameters.

Before evaluating automatic WML segmentation methods, reference WML segmentations annotated on 3T and ULF data were compared to assess the feasibility of WML detection and segmentation at ULF relative to clinical MRI. While prior studies have examined radiological detection of WMLs at ULF compared to 3T, this study is the first to compare reference annotated segmentation masks between ULF and 3T. While ULF and 3T manual lesion volumes were highly correlated, the mean DSC overlap score was only 0.30 ± 0.02. This improves on previous research which reported a DSC overlap score of 0.2 between ULF and 3T MIMoSA WML masks (Arnold et al., 2022). In the present study, while ULF overestimated total lesion volume, fewer individual lesions were detected compared to 3T, suggesting that the larger voxel size contributes primarily to volume overestimation. This is further supported by the analysis in a subset of lesions reviewed radiologically, where smaller lesions were more likely to be undetectable at ULF, highlighting the size-dependent limitations of ULF imaging. However, lesion location also appears to influence detection, as lesions may be missed in specific regions, such as the cerebellum or brainstem, where SNR is lower.

All PLAn models consistently demonstrated the highest overlap metrics with ULF manual segmentations and the strongest volumetric correlations with both 3T and ULF manual lesion masks. A total of four PLAn models were trained. PLAn 1 FLAIR and PLAn 1 FLAIR/T1w were pre-trained using the training cohort from the original PLAn publication (Donnay et al., 2023), down-sampled to $1.7 \times 1.7 \times 5$ mm resolution. The PLAn 2 models used the same participants from the paired 3T-ULF dataset, undergoing pre-training at 3T and fine-tuning at ULF. The segmentation labels used for training also differed between the models: PLAn 1 models were trained on C-DEF-generated WML masks that were not manually adjusted, whereas PLAn 2 models used manually corrected C-DEF WML masks. Surprisingly, this manual refinement of training labels did not improve the performance of PLAn 2 models over PLAn 1. This could reflect the fact that, once fine-tuned on the same ULF scans, both models converge toward similar ULF-specific features regardless of their upstream differences. Additionally, manual edits, mostly boundary

refinements, may have been diminished by downsampling the WML label to ULF resolution, providing little new information for the downstream task.

The nnU-Net models were more conservative in lesion segmentation compared to PLAn models. In the original publication on 7T scans, PLAn outperformed nnU-Net in detecting large, diffuse WMLs. When applied to the ULF segmentation task, pre-training PLAn on 3T data primarily enhanced the detection of lesions that were completely missed by nnU-Net. By first training on downsampled 3T images and labels, PLAn was likely able to learn more comprehensive and generalisable lesion features before fine-tuning on ULF data. This pre-training and fine-tuning approach helped PLAn better capture lesion variability and subtlety, whereas nnU-Net, trained only on the smaller ULF dataset, produced more cautious segmentations and missed more lesions. Additionally, PLAn models showed similar performance trends across different ULF acquisition parameters, which could be explained by the fact that PLAn was exposed to a broader range of lesion appearances and imaging characteristics during pre-training. In contrast, nnU-Net's performance was more variable and tended to favour cohorts with parameters overrepresented in the training dataset. This indicates that PLAn may be more generalisable across diverse ULF scans, while nnU-Net may be more susceptible to dataset biases. However, this conclusion is based on exploratory results where the dataset was split into unequal cohorts with varying scan parameters. Future studies should explore generalizability to larger datasets with varying acquisition parameters.

Despite the well-documented benefit of multi-contrast information for lesion segmentation at higher field strengths (Anbeek et al., 2004; Steenwijk et al., 2013), this study's results indicate that using both FLAIR and T1w at ULF did not improve segmentation performance for either nnU-Net or PLAn models. A key concern when incorporating ULF T1w was that inhomogeneous T1 hypointensities might reduce WML segmentation accuracy. However, the findings suggest this was not a major issue, as models trained with both contrasts performed similarly to those using only FLAIR. This is particularly useful given the extended ULF acquisition times, as it highlights that FLAIR alone can provide sufficient information for accurate segmentation without the need for additional scanning.

Automatic segmentation models trained on ULF data consistently outperformed MIMoSA, which was originally trained on 3T data. Among all methods, MIMoSA exhibited the lowest overlap metrics with ULF Aided masks

and significantly overestimated total WML volume compared to both 3T and ULF manual lesion masks. Additionally, MIMoSA's correlation with 3T and ULF lesion volumes was notably weaker ($\rho = 0.41$, Spearman) than that of other methods. These findings contrast with a previous study that found no significant differences in ULF and 3T lesion volumes and reported a stronger correlation ($r = 0.89$, Pearson) between ULF and 3T volumes. Several factors may explain this discrepancy. The prior study had a smaller sample size ($n = 33$) and used MIMoSA for both 3T and ULF segmentation, whereas in our study, 3T reference segmentations were manually refined from C-DEF outputs. Additionally, differences in cohort lesion burden and preprocessing methods, such as skull-stripping, may have influenced segmentation performance.

Similar to MIMoSA, SynthSeg also overestimated WML volumes compared to 3T and ULF manual masks. Although SynthSeg is trained on synthetic MRI datasets spanning multiple field strengths and resolutions, its training data include both vascular lesions and MS lesions. This may have contributed to the overestimation of WML segmentation around the ventricles, a region commonly affected by vascular lesions.

One limitation of this study is the use of MIMoSA, a segmentation model trained on 3T data, as a comparison for ULF segmentation performance. This choice was based on its use in a previous study (Arnold et al., 2022). Nevertheless, it is unsurprising that models trained on ULF data performed better than a segmentation model developed for high-field MRI. No attempt was made to adapt MIMoSA to ULF data within the present study, which could have been achieved through transfer learning or fine-tuning on ULF data. Training MIMoSA with ULF-specific lesion masks or applying techniques such as down-weighting false-positive regions (e.g., the sagittal sinus) could have improved its robustness across field strengths. However, enhancing MIMoSA's performance was beyond the scope of this study, and it was primarily included as a benchmark comparison.

Additionally, fine-tuning parameters for nnU-Net or PLAn models were not explored, which could have further optimised segmentation performance. In the original PLAn publication, a 7T model was fine-tuned with only five cases in the training dataset (Donnay et al., 2023). Given the lower resolution and larger voxel sizes at ULF, 20 participants were chosen for fine-tuning in this study, as a smaller training set could have led to limitations in the diversity and

number of image patches extracted for training. This can reduce the variability seen during model training, increasing the risk of overfitting and reducing generalizability to new data. However, Systematic evaluation of how changes in the pre-training or fine-tuning datasets influence the final segmentation results was not conducted. Future studies should explore this issue further, as manual annotation for training datasets is time-consuming, and dataset size can impact overfitting risks, especially in DL-based segmentation models.

Another key limitation is the lack of generalizability testing for the models. Hyperfine ULF MRI data is relatively new, and there are no openly available, large-scale MS lesion datasets for ULF MRI to independently assess generalizability. Future work should aim to evaluate nnU-Net and PLAn segmentation models on external datasets or incorporate more diverse training data to improve model robustness. Importantly, PLAn and nnU-Net models were trained on our dataset and reference WML masks, which may introduce bias towards the criteria adopted in our lab for manual labelling. In contrast, MIMoSA and SynthSeg were not exposed to any of our manual masks prior to validation. However, the purpose of MIMoSA and SynthSeg in this study was to provide benchmarks of what tools are currently available for ULF segmentations.

The ULF dataset used in this study was heterogeneous in acquisition parameters, which may have influenced segmentation outcomes. Given the retrospective nature of this study, all available cross-sectional data were included, regardless of differences in software versions, reconstruction methods, or voxel sizes. In a preliminary investigation, the impact of imaging acquisition parameters on segmentation quality was assessed by grouping scans based on software versions, reconstruction methods, and voxel sizes. These parameters were chosen based on radiologist input, as anecdotal evidence suggested that certain software updates produced poorer FLAIR image quality for MS lesion detection. Additionally, different voxel sizes introduce varying degrees of partial volume effects and may miss small lesions, potentially affecting overlap metrics and volume estimates. However, because evenly distributed or controlled groups across different acquisition parameters were not available, direct comparisons of their impact on segmentation performance could not be performed. Unequal group sizes and differences in total lesion burden across subgroups may have influenced segmentation accuracy, making it difficult to draw definitive conclusions about the role of acquisition parameters in ULF lesion detection. Instead, the evaluation focused on whether general trends in

segmentation performance across different methods remained consistent, rather than making direct comparisons between acquisition settings. Future studies should take a more systematic approach to assessing the influence of acquisition parameters on WML detection and segmentation quality. This could involve prospective studies or post-acquisition image reconstruction in collaboration with Hyperfine, allowing for comparisons between DL reconstruction and conventional Cartesian sampling.

Finally, acquisition parameters were not considered when creating the training dataset, despite their potential impact on model performance. If acquisition settings differ significantly between training and validation cohorts, segmentation accuracy may be affected. Future studies should investigate whether training models on more diverse acquisition settings could improve their generalizability across different ULF scans.

Overall, the findings support the feasibility of automated WML segmentation at ULF, highlighting its potential for low-cost, portable, and scalable MS monitoring.

4.2 Subtraction Imaging Pipeline to Assess Longitudinal Lesion Changes in Multiple Sclerosis at Ultra-Low-Field

4.2.1 Introduction

An exciting prospect of portable ULF MRI is its promise to enable cost-effective and frequent longitudinal follow-up scans at the point of care, which is particularly relevant in neurological disorders, including MS. In clinical practice, the radiological evaluation of new lesions and changes in existing lesions helps monitor disease progression and guides therapeutic decisions (Wattjes et al., 2021). Traditionally, longitudinal MRIs are assessed using side-by-side comparisons of scans, with each timepoint displayed in multiple planes across adjacent viewports. This approach can be time-consuming, particularly in patients with a high lesion burden, as each lesion is evaluated individually. Furthermore, MRI scans may not be properly registered in clinical settings, as images are usually displayed as acquired directly from the scanner, hence not consistently oriented or sliced between timepoints. This potential lack of alignment can further lead to inaccuracies and errors in radiological lesion assessment.

To address these challenges, voxel-wise subtractions of registered and normalised longitudinal MRI images have emerged as a valuable tool in the radiological assessment of MS lesions (Eichinger et al., 2019; Lee et al., 1998). Subtraction imaging aims to accentuate the visualisation of structural changes, such as the emergence of new lesions, while reducing the visibility of stable lesions that would be cancelled out during subtraction. Previous studies have demonstrated that, at clinical field strengths, subtraction maps significantly improve the detection of new lesions or lesion changes, reduce the time required for radiological assessments, and enhance inter- and intra-rater agreement of lesional burden (Eichinger et al., 2019; Horsfield et al., 2016; Patel et al., 2017; Tan, van Schijndel, Fazekas, Filippi, Freitag, Miller, Yousry, Pouwels, Ader and Barkhof, 2002). Subtraction pipelines show sensitivity to subtle changes in MS lesion burden and increase power in clinical trials (Moraal et al., 2009, 2010).

Three fundamental pre-processing steps prior to subtraction are co-registration, bias-field correction, and intensity normalisation (Im et al., 2024). Image registration, which aligns images to the same spatial coordinates and resolution, is necessary for voxel-wise subtraction. Typically, rigid registration is applied, ensuring that the structural integrity of key features is maintained by aligning the images without distorting their inherent shapes (Smith et al., 2007). Either follow-up images are aligned to baseline scans, or all images are registered to a common neutral space. Registering all images into a common neutral space ensures that the same processing steps, including interpolation, are applied to all scans, not just the follow-up images. This uniform processing helps maintain consistency across the images, reducing potential biases or variations that could arise if only the follow-up scans were interpolated. However, it necessitates re-processing each time a new point is added.

Bias-field correction and intensity normalisation allow consistent intensity values within and across different scans. Bias-field correction mitigates the influence of low-frequency intensity variations caused by magnetic field inhomogeneities during acquisition (Tustison et al., 2010). Additionally, MRI scans are typically acquired in arbitrary units that are not directly comparable across different study visits, even within a single subject. Intensity normalisation standardises MRI intensities and addresses variations in acquisition parameters between scans. For subtraction imaging, tissues of interest across longitudinal within-subject scans should have comparable intensities to allow for longitudinal scans. While simple intensity scaling techniques can be used across similar

modalities and images, more advanced methods have been developed to account for heterogeneous brain scans. One such method is *WhiteStripe*, a linear, tissue-specific intensity normalisation technique that identifies NAWM in the brain MRI and uses its intensity values to scale the entire image (Shinohara et al., 2014). By standardizing based on NAWM, *WhiteStripe* effectively reduces variability caused by differences in scanner types or acquisition protocols, improving consistency across longitudinal scans.

Subtraction pipelines could be a valuable tool for radiological evaluations at ULF. However, generating ULF subtraction maps presents additional challenges, primarily due to the lower image quality and the fact that existing open-source MRI software tools for performing the main pre-processing steps are optimised for higher-field MRI systems. While prior research has highlighted the clinical utility of ULF MRI in detecting MS lesions (Arnold et al., 2022; Okar et al., 2025), the sensitivity of detecting longitudinal changes of MS lesions at ULF is not known. In this section, a ULF subtraction pipeline was developed and using the resulting ULF subtraction images, the ability to radiologically detect new, enlarging, and shrinking MS lesions on ULF MRI compared to 3T was assessed. Additionally, the minimum extent of lesion change detectable at ULF and the factors influencing the detectability of such changes were explored.

4.2.2 Methods

MRI data were collected for 23 participants with MS or probable MS at two timepoints (mean follow-up interval = 16.9 ± 1.86 months) as part of a natural history protocol (Appendix A.1). Paired same-day HF (3T) and ULF (64mT) MRI scans, including FLAIR contrasts, were acquired at both timepoints for all participants. Full dataset details, including acquisition parameters and participant characteristics, are provided in Appendix A.4.2.

4.2.2.1 Overview of Subtraction Pipelines

An overview of the study design is provided in Figure 4.9A. Paired, same-day 3T-ULF FLAIR scans were processed separately using specific pipelines to produce preprocessed longitudinal MRI scans and subtraction maps (SUBLONG). The pipelines were optimised on a subset of 5-10 scans by testing various tool combinations and having a neurologist with 7 years of experience in neuroimaging (S.V.O.) qualitatively evaluate the pipeline outputs (data

not shown). Pipelines were tested both with and without skull-stripping before registration (*BET*, *SynthStrip*), with different registration tools (AFNI's *3dAllineate*, *FLIRT*, *antsRegistration*), and under varying registration parameters (interpolation method, degrees of freedom). Additionally, several intensity normalisation methods were tested, including the routine median-based normalisation used in the TNS lab's clinical 3T SUBLONG pipeline, and more advanced methods like the *WhiteStripe* method (also used in MIMoSA's preprocessing pipeline in Chapter 4.1. Given the inherent differences in image characteristics between ULF and 3T, the same open-source tools and settings did not always produce optimal results across both field strengths, resulting in field-strength-specific but functionally equivalent pipelines.

Figure 4.9B outlines the image processing steps in our ULF subtraction pipeline. First, ULF FLAIR images were N4 corrected and resampled to isotropic resolution ($1.7 \times 1.7 \times 1.7$) using AFNI's *3dresampling* tool with linear interpolation (Cox, 1996; Tustison et al., 2010). All ULF images were then skull-stripped using *SynthStrip* (Hoopes et al., 2022). A participant-specific template was created using the ANTs multivariate template construction tool, initialised with the baseline scan (Avants et al., 2011). All ULF images were linearly registered (12 degrees of freedom) to the template scan. Although subtraction imaging typically relies on rigid registration, during pipeline optimization it was found that affine registration more effectively aligned ULF scans. This could be attributed to the larger voxel sizes in ULF imaging, which introduce more geometric differences between scans. The added flexibility of affine registration may have allowed for better adjustment of these differences, leading to more accurate alignment. Finally, skull-stripped and registered images were intensity-normalised using *WhiteStripe* normalisation (Shinohara et al., 2014) to produce *ULF preprocessed* images. Given potential variability in ULF acquisition parameters or hardware/software versions between timepoints, *WhiteStripe* normalisation was chosen as it more effectively produced consistent intensity profiles across longitudinal scans than other methods (see Appendix D.2.1 for an example). ULF FLAIR follow-up images were subtracted from ULF baseline to create voxel-wise subtraction maps, referred to as *ULF SUBLONG*.

All 3T FLAIR images were N4 bias-field corrected (Hoopes et al., 2022) and skull-stripped with *SynthStrip* (Jenkinson and Smith, 2001). Next, follow-up images were rigidly registered with FSL's *FLIRT* using trilinear interpolation

(Jenkinson et al., 2002; Jenkinson and Smith, 2001). All images were normalised using median-based normalisation to create *3T preprocessed* images. In this normalisation, the median value of the image intensities is computed, and the image is scaled such that its median matches a reference value of 500. Median-based normalisation was used for 3T images because the 3T FLAIR protocol is more stable across timepoints than the ULF FLAIR, and this approach is standard in the TNS lab's clinical 3T SUBLONG pipeline. 3T subtraction maps, or 3T SUBLONG, were created by performing a voxel-wise subtraction of the follow-up images from the baseline. Finally, the 3T preprocessed and SUBLONG maps were rigidly registered to ULF while preserving the original 3T resolution. This allowed for easy cross-referencing of radiological readings across field strengths without compromising 3T resolution. For 3T follow-up images that were first registered to the 3T baseline, the registration matrices between the 3T follow-up to baseline and then baseline 3T to ULF were concatenated to reduce interpolation steps.

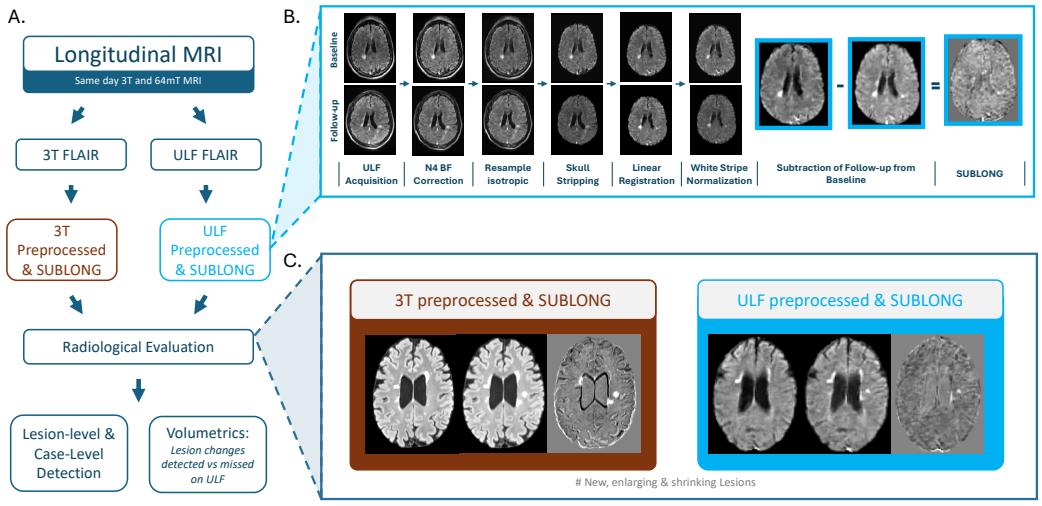


Figure 4.9: Overview of study design. A) Paired same-day ULF and 3T FLAIR images were acquired, preprocessed, and evaluated to assess longitudinal lesion changes detectable on ULF. B) Image processing pipeline developed to create ULF preprocessed and ULF SUBLONG images. C) Tandem 3T-ULF radiological evaluation of longitudinal changes in lesions.

4.2.2.2 Radiological Assessment of Lesion Evolution

A neurologist with 7 years of neuroimaging experience (S.V.O.), blinded to clinical and demographic information, assessed the longitudinal evolution in MS lesions. The neurologist performed side-by-side longitudinal evaluations of preprocessed ULF images, ULF SUBLONG, preprocessed 3T images, and 3T

SUBLONG using *3D Slicer* (https://www.slicer.org/). ULF and 3T radiological assessments were done in tandem to establish a foundational understanding of lesion detectability at ULF under optimal conditions, as longitudinal lesion evolution in MS has, to our knowledge, never been investigated at ULF.

For each participant, S.V.O. identified new (NL) and changing lesions (CL) based on predefined criteria (Figure 4.10). CLs include both enlarging (EL) and shrinking lesions (SL). The predefined criteria include guidelines for identifying lesion dynamics using the preprocessed longitudinal images and SUBLONG. These criteria include guidelines for assessing lesion dynamics using preprocessed longitudinal images and the SUBLONG sequence. For instance, an SL is characterized by a visible reduction in size on the follow-up preprocessed image compared to the baseline, accompanied by a hypointense halo on the SUBLONG image, indicating the area of the lesion that was present at the first timepoint but not at the second (Figure 4.10, first row).

An important consideration in identifying EL is the potential effect of ventricular enlargement, which can cause lesions to appear displaced, creating a halo-like artifact on the SUBLONG image similar to what would be expected in an EL. This effect is less problematic for lesions located centrally in the WM, where false positive enlargement would typically appear as opposing bright and dark edges. Given that ventricular enlargement in MS averages approximately 2.4% per year and the mean follow-up time in this study was 1.4 years (Sinnecker et al., 2020), ventricular expansion is likely undetectable for many participants in this cohort. Nevertheless, ventricle expansion, manifesting as a dark halo around the ventricles in the SUBLONG images, was carefully considered when determining EL status, especially for periventricular lesions.

All ULF radiological readings were validated using the 3T radiological readings as a reference to assess what radiological lesion dynamics can be detected at ULF compared to a 3T clinical standard. Each NL or CL identified at ULF was categorised as a true positive (TP) or a false negative (FN). Since ULF radiological assessments were identified in tandem with the 3T reference, false positives were not applicable. Lesion-level true negatives (TN) were not identified due to time constraints. Case-level TP, FN, and TN were computed as changes detected at the case level (i.e., whether any new lesions were detected in a subject). Case-level measurements are clinically useful because they capture whether a case is experiencing overall disease progression. A total number of case-level TPs, TNs, and FNs were reported and case-level sensitivity and

positive predictive value (PPV) and negative predictive value (NPV) for CLs and NLs were calculated. At the lesion level, the total number of TPs and FNs were reported and lesion-level sensitivity for CLs and NLs was calculated.

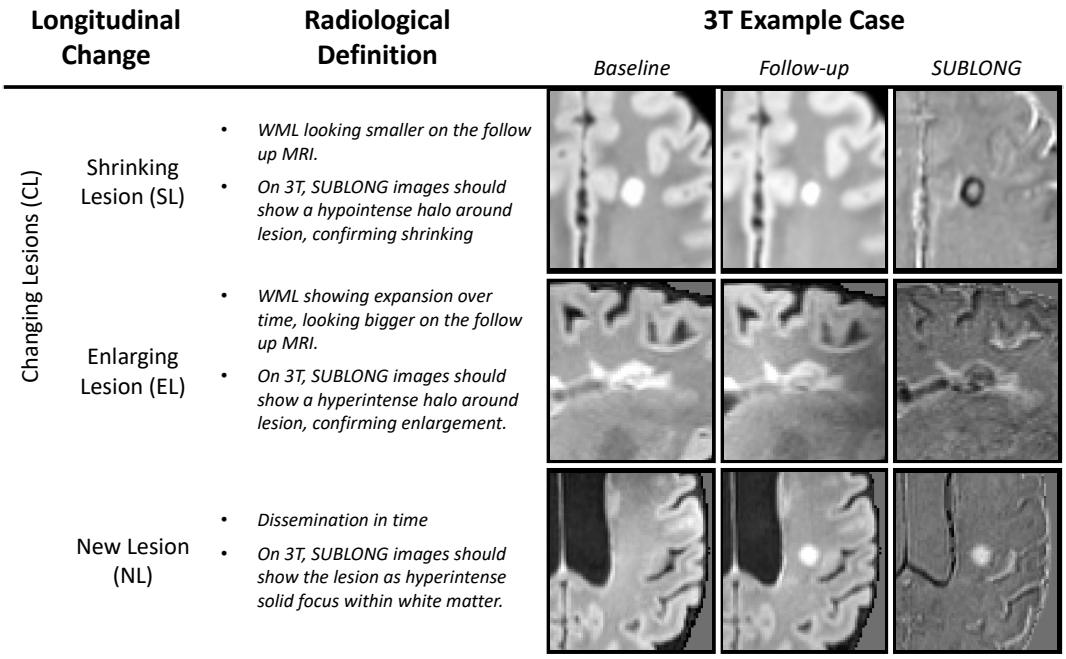


Figure 4.10: Radiological definitions of changing lesions (CL) including enlarging (EL) and shrinking lesions (SL) and new lesions (NL) using preprocessed and SUBLONG images. A visual example is shown for each definition on preprocessed 3T images and 3T SUBLONG.

4.2.2.3 Quantitative Analysis of Lesion Evolution

Radiologically determined CLs and NLs were manually segmented on the preprocessed images at 3T and ULF. CLs were segmented at baseline and follow-up. NLs were segmented on follow-up images. 3T volumes of lesions and 3T longitudinal differences in volumes were compared for radiological assessments captured at ULF and undetected at ULF. For radiological assessments captured at ULF, 3T and ULF volumes of lesions and longitudinal differences in lesion volume were quantified and compared.

4.2.2.4 Analysis of Results across ULF Acquisition Parameters

In a preliminary analysis similar to the one performed in the previous section (Chapter 4.1.2.7), the influence of ULF acquisition-related metrics on radiological and quantitative lesion dynamics results was qualitatively examined. CL radiological detection, NL radiological detection, and differences between 3T

and ULF longitudinal CL volume changes were assessed against heterogeneity of baseline and follow-up ULF reconstruction method, in-plane voxel size, and software version. The difference between 3T and ULF NL volume was assessed against the follow-up ULF reconstruction method, in-plane voxel size, and software version. While quantification was not possible due to the small sample size, trends were visually assessed to identify potential influences of ULF acquisition software on the results.

4.2.2.5 Statistical Analysis

For case- and lesion-level radiological assessments, sensitivity, NPV, and PPV were defined as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}; \quad \text{NPV} = \frac{TN}{TN + FN}; \quad \text{PPV} = \frac{TP}{TP + FP}$$

All lesion volume data were assessed for normality using the Shapiro-Wilk test. For the longitudinal analysis of CLs (analyses combining EL and SL), the longitudinal volume change was calculated as the absolute volume difference between the follow-up and baseline or as the absolute percent change from the baseline. This approach allowed the analysis of lesion size evolution irrespective of direction, enabling grouped assessment of enlarging and shrinking lesions.

To compare 3T volumetric measures of CLs and NLs that were radiologically detected on ULF versus those that were not, either a two-tailed unpaired t-test with Welch's correction (when normality was confirmed) or a two-tailed unpaired Mann-Whitney U test (for non-normal distributions) was used. Welch's correction was applied to account for unequal group sizes and the possibility of unequal variances.

For within-lesion comparisons of lesion volume at baseline and follow-up (at either 3T or ULF), the two-tailed paired Wilcoxon signed-rank test was used. Although ELs were expected to increase, a two-tailed test was chosen to provide a more conservative estimate, given the possibility of volumetric disagreement with radiological assessments.

Correlations between ULF and 3T volumetric measurements were assessed using Spearman's rank correlation coefficient (ρ). Bland-Altman analysis was used to evaluate agreement between ULF and 3T volume measurements. For CLs, Bland-Altman plots were generated based on the absolute percent volume change from baseline to follow-up (difference plotted against the mean of ULF and 3T values). For NLs, Bland-Altman analysis was used to assess raw

volume differences between ULF and 3T. Bias was calculated as the mean difference between methods, and limits of agreement (LoA) were defined as ±1.96 standard deviations of the difference.

4.2.3 Results

4.2.3.1 Radiological Assessment of Lesion Evolution

Across all participants ($n = 23$), fourteen had at least one SL, EL, or NL, based on the 3T reference radiological evaluation (Table 4.6). SL, EL, and NL cases were visible on the ULF preprocessed and SUBLONG images. SUBLONG radiological characteristics, as described in Figure 4.10, were visible on ULF SUBLONG. Figure 4.11A shows an example of an SL visible on 3T and ULF with a hypointense halo on the SUBLONG images and an example of an EL with a hyperintense halo confirming lesion enlargement (Figure 4.11B).

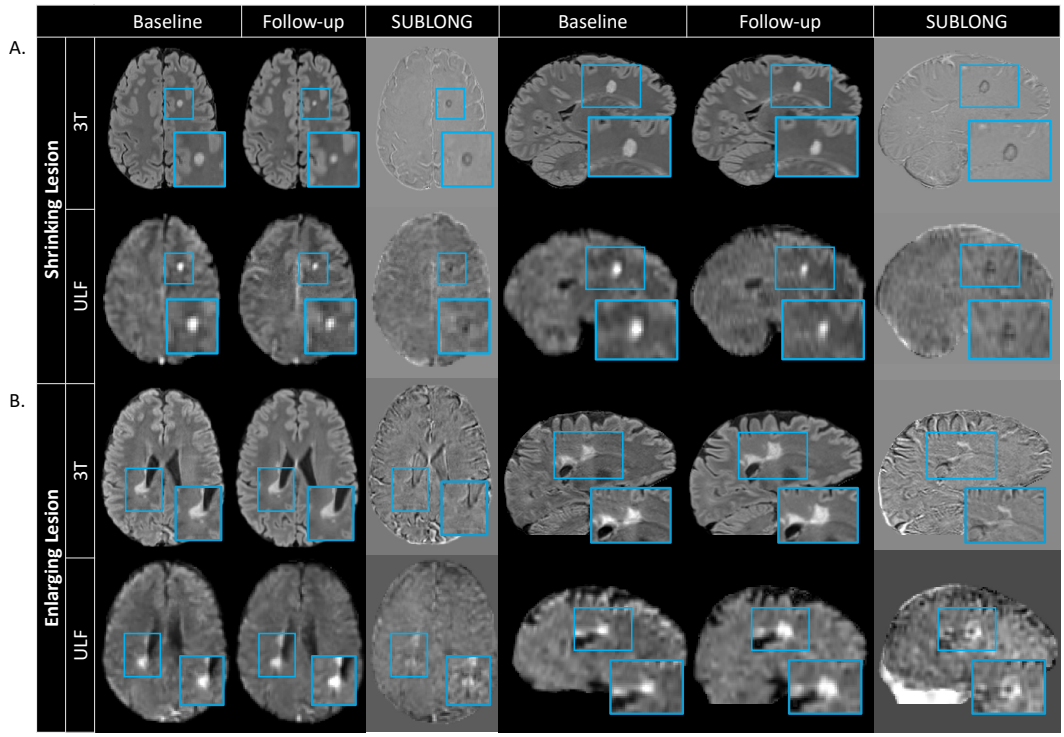


Figure 4.11: Examples of CLs visible on both 3T and ULF across preprocessed and SUBLONG images. A) Axial and sagittal whole-brain FLAIR images at baseline and follow-up showing a shrinking lesion, visible at both 3T and ULF, with a hypointense halo on SUBLONG. B) Axial and sagittal whole-brain FLAIR images at baseline and follow-up showing an enlarging lesion, also visible at 3T and ULF, with a hyperintense halo on SUBLONG.

In total, 23 CL (SL+EL) were identified across 11 participants. At least one CL was visible on ULF in all 11 participants, resulting in a case-level sensitivity and NPV of 100%. At the lesion level, 18 out of 23 CL were visible on ULF,

corresponding to a lesion-level sensitivity of 78%. While all three SL were visible as shrinking on ULF, 5 of the 20 ELs were not visible as enlarging on ULF. Across all participants with PV EL, only one participant demonstrated clear ventricular enlargement. Two additional participants showed possible subtle ventricular expansion, each with one PV EL. Notably, all three of these PV ELs were detected as enlarging on ULF. This suggests that while ELs that are confounded by ventricular expansion might be easier to detect on ULF, ventricular enlargement was not the primary driver of our results; most ELs detected on ULF were not influenced by ventricular enlargement.

At least one NL was identified on ULF in 3 participants, yielding a case-level sensitivity of 60% and NPV of 90%. Noise in the WM made it difficult to confirm small NLs on the ULF SUBLONG. At the lesion level, only 5 of the 14 NL identified on 3T were also visible on ULF, corresponding to a lesion-level sensitivity of 35.7% (Table 4.6).

Table 4.6: Longitudinal Lesional Change at Case and Lesion Level

Longitudinal Lesional Change	Case Level	Lesion Level
Changing Lesions (Shrinking + Enlarging)		
Total (3T Reference)	11	23
# TP	11	18
# TN	12	NA
# FN	0	5
Sensitivity	100%	78%
NPV	100%	NA
Shrinking lesion		
Total Lesions (3T Reference)	3	3
# TP	3	3
# TN	20	NA
# FN	0	0
Sensitivity	100%	3/3, 100%
NPV	100%	NA
Enlarging lesions		
Total (3T Reference)	9	20
# TP	8	15
# TN	14	NA
# FN	1	5
Sensitivity	89%	75%
NPV	93%	NA
New lesions		
Total (3T Reference)	5	14
# TP	3	5
# TN	18	NA
# FN	2	9
Sensitivity	60%	35.7%
NPV	90%	NA

Abbreviations: TP = true positive; TN = true negative; FN = false negative; NPV = negative predictive value;

4.2.3.2 Quantitative Analysis of Lesion Evolution

3T-derived lesion volume measures for CLs and NLs for cases whose longitudinal evolution was radiologically detectable on ULF (detected on ULF) were compared to cases whose longitudinal evolution was not detected on ULF (undetected on ULF).

Across all CLs (EL+SL), the 18 lesions detected on ULF did not have significantly higher baseline 3T volume ($434 \pm 82.5\ mm^3$) compared to the five lesions that were undetected on ULF ($369.2 \pm 168\ mm^3$; $U = 35$, $p = 0.50$;

Figure 4.12A). However, the absolute change in 3T lesion volume from baseline to follow-up was significantly greater for the ULF-detected CLs (36.2 ± 14.48 mm^3) compared to ULF-undetected CLs (12.0 ± 30.14 mm^3; $t(21) = 2.7$, $p = 0.016$; Figure 4.12B), but not when comparing the absolute percent volume change ($U = 23$, $p = 0.11$; Figure 4.12C). Overall, 74% of CLs were detected when the 3T absolute percent change was $> 9\%$.

NLs detected at ULF had significantly larger volume on 3T (331 ± 83.6 mm^3) than those not detected at ULF (93.9 ± 35.8 mm^3; $U = 4$, $p = 0.012$; Figure 4.12D). The smallest NL detected on ULF was 113 mm^3, as measured on 3T. Although in the same participant, two larger NLs (248 mm^3 and 127 mm^3) were not detectable on ULF. Figure 4.12E shows an example of NLs that were captured on ULF and 3T, with a solid hyperintense area in 3 locations on the SUBLONG where the NL formed, and an example where two small NLs were not captured on the ULF preprocessed or SUBLONG.

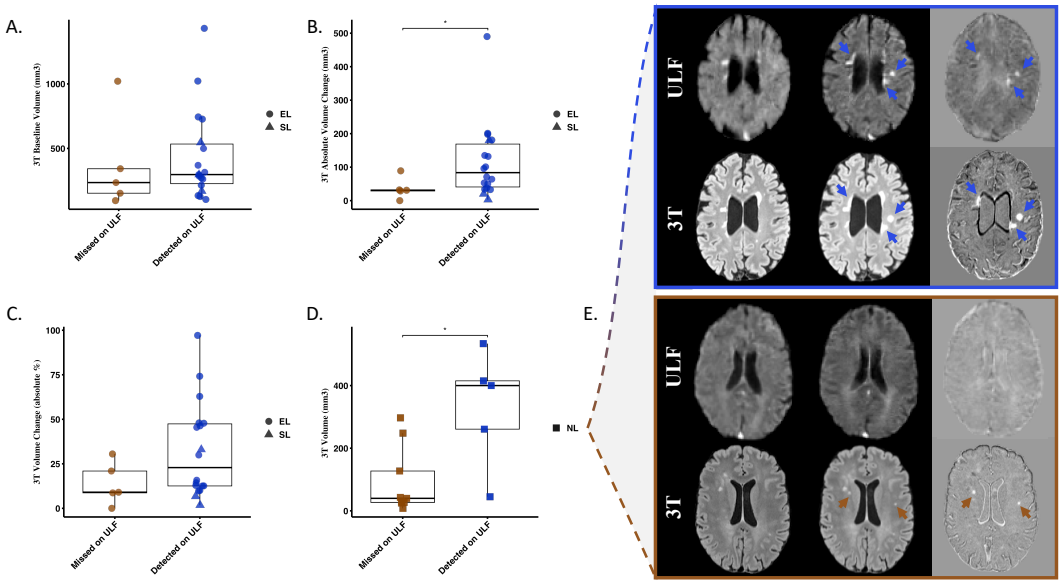


Figure 4.12: Quantitative analysis of radiologically identified lesion dynamics that were either detected (blue) or undetected (orange) on ULF. A) Baseline 3T lesion volumes for CLs (EL+SL), B) Absolute change in 3T lesion volume from baseline to follow-up in CLs (EL+SL), C) Absolute percent 3T volume change from baseline to follow-up for CLs (EL + SL), and D) 3T volumes of NL for lesion changes detected vs. undetected on ULF. E) Example axial FLAIR images (baseline, follow-up, and SUBLONG) from ULF and 3T in a participant with NLs. The top panel (blue outline) shows a case where three new lesions were visible on both 3T and ULF (highlighted with arrows). The bottom panel (orange outline) shows a case where two new lesions were visible on 3T but not on the ULF images.

Next, 3T and ULF lesion volumes were compared for CLs or NLs detected on ULF. EL volume increased significantly over time when measured at both 3T ($W = 82.0$, $Z = 1.25$, $p = 0.018$) and ULF ($W = 116.0$, $Z = 3.18$, $p < 0.001$; Figure 4.13A). These results remained significant even after removing

one outlier with a disproportionately large follow-up volume at ULF compared to 3T. After excluding this case, EL volume still showed a significant increase over time at both 3T ($W = 73.0$, $p = 0.020$) and ULF ($W = 101.0$, $p < 0.001$). Due to the small number of SL ($n = 3$), statistical comparisons were not conducted. Qualitatively, only one SL showed a substantial 3T volume decrease (-180 mm^3), while the other two were stable or slightly increased (2% and 7%). On ULF, two of these SLs showed a decrease in volume, and one remained unchanged (Figure 4.13B).

Across all CLs, the difference in lesion volume from baseline to follow-up measured at ULF and 3T was significantly correlated ($r(18) = 0.61$, $p = 0.007$; Figure 4.13C). However, ULF measurements showed significantly higher absolute percent volume change than 3T (ULF: $101.1 \pm 38.6\%$, 3T: $32.3 \pm 6.3\%$; $W = 135$, $p = 0.0019$). Bland-Altman analysis revealed a mean bias of 68%, indicating a systematic overestimation of percent volume change at ULF compared to 3T (Figure 4.13D).

These findings remained consistent after removing the same outlier described above. The change in EL volumes measured at 3T and ULF remained significantly correlated ($r(17) = 0.70$, $p = 0.003$). While ULF still overestimated the percent volume change, the overestimation was slightly reduced compared to the full dataset. ULF percent volume change remained significantly higher than 3T (ULF: $80.4 \pm 21.0\%$, 3T: $33.3 \pm 6.6\%$; $W = 117$, $p = 0.003$). Bland-Altman analysis without the outlier showed a mean bias of 62.9% ($\pm$87.6%), with LoA ranging from -108% to 234.7%.

For the five NLs visible on ULF, 3T and ULF lesion volumes were not significantly correlated ($r(5) = 0.49$, $p = 0.40$; Figure 4.13E). In four of five cases, ULF-derived lesion volumes were larger than the corresponding 3T volumes. Bland-Altman analysis showed an average bias of +120 mm^3 (Figure 4.13F), suggesting an overestimation of NL volume at ULF.

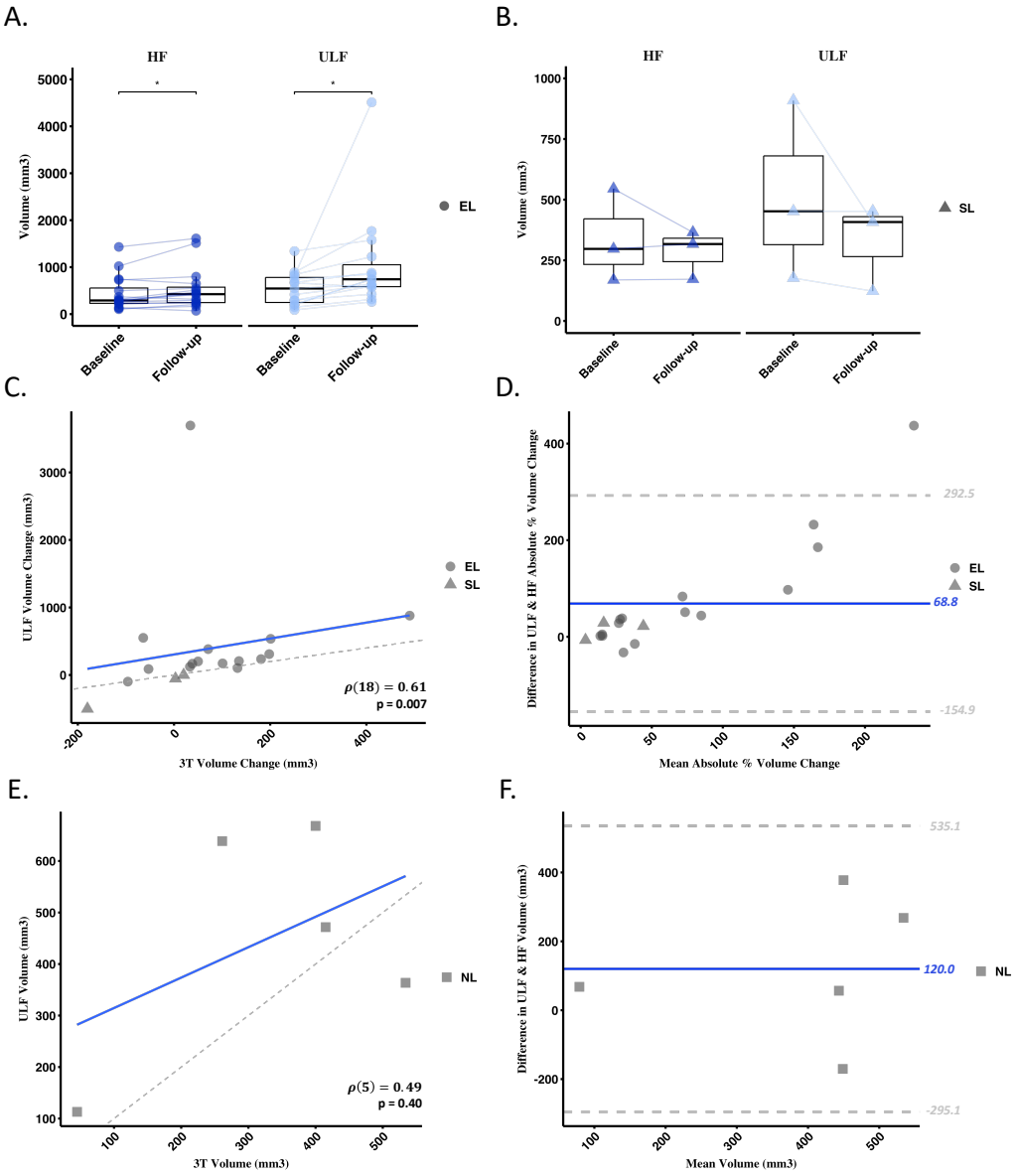


Figure 4.13: Comparison of 3T and ULF volume measures for EL, SL and NL radiologically detected on ULF. A) EL show increased volume changes at follow-up compared to baseline at 3T and ULF. B) SL volume at baseline and follow-up as measured by 3T and ULF. C) 3T and ULF volume changes (follow-up – baseline) were significantly correlated. However, correlation line (blue) is consistently higher than X=Y (dashed light grey line), showing higher volume differences at ULF than 3T. D) Bland-Altman plot showing the level of agreement between 3T and ULF absolute percent change in volume of CL (dark blue line bias, limits of agreement are light grey dashed lines). E) The 5 3T and ULF NL volumes measured on follow-up images are not significantly correlated (blue line). ULF overestimated most of the NL volume compared to 3T. F) Bland-Altman plot showing the level of agreement between 3T and ULF NL volume (dark blue line bias, limits of agreement are light grey dashed lines).

4.2.3.3 Analysis of Results across ULF Acquisition Parameters

Given the heterogeneity in ULF acquisition parameters, the impact of factors such as reconstruction method, voxel size, and scanner software version

on radiological detection of lesion evolution on ULF was assessed (full results are available in Appendix D.2.2). However, there was too much variability in acquisition parameters across participants to determine whether and how acquisition parameters affect SUBLONG image quality. In all participants except one, there was a software version change between baseline and follow-up timepoints, resulting in a change in in-plane voxel size between baseline and follow-up (ranged between 1.6–1.8 mm^3) (Appendix D.2.2). Ten participants had baseline scans with DL reconstruction method and follow-up scans with Cartesian reconstruction methods. One participant had both timepoints DL reconstruction, and the rest had Cartesian reconstruction for both timepoints.

Participants with the highest number of undetected CLs often had either a change in reconstruction method between timepoints or scans acquired with software version *8.3.2*, which is known to degrade FLAIR image quality (Appendix D.2.2). However, lesion evolution was still detectable in other cases with similar parameters, and no consistent association was observed across the dataset.

4.2.4 Discussion

In this section, I present a longitudinal subtraction pipeline to support the radiological detection of lesion dynamics, which was then used to assess which types of longitudinal lesion changes could be identified on ULF MRI compared to 3T clinical standard images. The findings demonstrate that key radiological features of EL, SL, and NLs can be identified on ULF SUBLONG images to some extent, highlighting the potential value of longitudinal subtraction techniques at ULF. In all participants with ELs or SLs (collectively referred to as CLs), at least one CL was detected on ULF, resulting in a case-level sensitivity of 100%. In contrast, NL detection was more limited and challenging, with a case-level sensitivity of 60%. The absence of visible changes and the low detection of NL can be attributed to several factors. First, new lesions were uncommon in our cohort, with only three cases exhibiting NL, likely due to the effectiveness of disease-modifying therapies. Among these, the detection of NL was volume-dependent, aligning with previous literature indicating that MS lesions are typically visible only when larger than 4 mm in diameter (Arnold et al., 2022). Most of the NL observed were smaller than this threshold. Lesion dynamics that were undetected on ULF were smaller in volume (for NLs),

exhibited lower absolute volume change (for CLs), or showed more diffuse signal on 3T. Moreover, ULF-derived lesion volumes tended to be overestimated: CLs exhibited a significantly greater percent volume change on ULF compared to 3T, and NL volumes were consistently larger at ULF. Variations in scan protocols and software versions created significant challenges for key image processing steps in this study, as discussed below. These findings highlight both the promise and current limitations of using ULF MRI for longitudinal lesion assessment in MS.

The quality of ULF SUBLONG maps varied significantly across participants, largely due to heterogeneous scan protocols and software versions. This variability complicated our ability to systematically assess the impact of specific acquisition parameters on subtraction image quality. High levels of WM noise further limited the reliability of detecting subtle lesion dynamics, as noise artifacts sometimes appeared as false-positive or false-negative changes on ULF SUBLONG images. While smoothing techniques may help mitigate this issue, image noise remains a major limitation of current ULF subtraction-based approaches. Standardizing ULF acquisition protocols and software versions will be essential for enabling meaningful comparisons and improving the consistency of radiological assessments in future studies.

At the same time, rapid technological improvements, including the introduction of new ULF scanners and updated software, mean that the limitations observed in this dataset may not apply to future clinical practice. While this evolution holds promise for improving image quality and lesion detectability, it also creates challenges for interpreting historical and follow-up data and assessing the generalisability of findings across different ULF platforms. To address these challenges, image harmonisation techniques such as *ComBat* have been successfully applied in multi-site and multi-scanner MRI studies to mitigate scanner-related variability while preserving biological signals (Fortin et al., 2018, 2016). Applying such methods to ULF datasets could facilitate combining earlier and newer data, improving comparability across evolving platforms. Going forward, it will be important to balance the need for stable, comparable datasets with the adoption of newer technologies that can push the boundaries of what is feasible at ULF (De Stefano et al., 2022).

Radiological ratings were performed in tandem on 3T and ULF images to determine whether longitudinal lesion dynamics apparent at a clinical standard (3T) could also be detected at ULF. To our knowledge, longitudinal MS

lesion evolution has not previously been analysed at ULF. These findings lay the groundwork for studies in which ULF ratings are performed independently by multiple neurologists and/or neuroradiologists: either fully blinded to 3T to mimic settings where 3T is not available or with access to 3T baseline alongside ULF baseline and follow-up scans to simulate realistic clinical workflows in which 3T is used at baseline but ULF is used for follow-up. Such future studies will be essential for determining the practical utility of ULF MRI as a complementary imaging tool in MS. Relevant to the fact that ULF images were assessed in tandem with 3T images, the number of lesion-level TNs was not determined in this study. To determine lesion-based TN, not only would all stable lesions need to be counted and identified at both ULF and HF, but each stable lesion would need to be cross-referenced between both scans. This process would be particularly tedious and time-consuming, especially in cases with high lesion loads. Automated segmentation algorithms, such as PLAn (Chapter 4.1), will help streamline such analyses. Nevertheless, case-level TN was measured, which is more relevant in clinical practice and still provides valuable insights for evaluating lesion progression.

Previous studies have shown that additional ULF acquisitions within the same session can improve the detectability of WML in MS. Image averaging across multiple acquisitions or across orthogonal acquisitions (e.g., one image with higher resolution in the axial plane with one with higher resolution in the coronal plane) can improve the visibility of WML and increase lesion conspicuity. In one study, a lesion not detectable on a single ULF axial FLAIR acquisition became visible after averaging multiple ULF FLAIR axial acquisitions (Arnold et al., 2022). Another study, where raters identified lesions on ULF while blinded to 3T, found that raters marked fewer false-positive areas when provided an additional plane along with a ULF axial FLAIR (Okar et al., 2025). Despite this prior evidence, our study did not assess the impact of multiplane or multi-acquisition averaging on radiological detection of WML change. Given that anisotropic ULF scan times are already comparable to typical 3T clinical scans, maximizing the diagnostic value of a single scan could improve scalability and cost-effectiveness by reducing the need for repeated acquisitions. Moreover, while additional FLAIR planes were collected in our protocol when time allowed, there were not enough cases with additional FLAIR planes at both baseline and follow-up to fully evaluate their impact. However, in the

next section, I explore how additional planes could be combined using super-resolution techniques to enhance WML visibility.

In summary, this study demonstrates the feasibility of using longitudinal subtraction imaging at ULF to detect evolving lesion dynamics in MS and explore the determinants of detectable changes at ULF. While limitations remain, most notably noisy WM and protocol variability, these findings establish a foundation for the potential ULF-based longitudinal lesion assessment in MS. Future work should prioritize protocol standardisation, explore enhanced interpolation and denoising techniques, and evaluate the added value of multi-plane or multi-acquisition approaches to improve sensitivity to lesional changes.

4.3 Improving Ultra-Low-Field MRI Resolution using Partially Acquired K-Space

4.3.1 Introduction

The fundamental limitation of ULF MRI is the reduced image quality, particularly the decreased contrast-to-noise and signal-to-noise ratios (CNR and SNR) associated with the lower magnetic field strength. The low SNR can be partially mitigated by averaging multiple acquisitions or by reducing the overall image resolution. However, reducing image resolution can also diminish the clinical utility of the MRI images, as it can result in missing small but clinically significant features, such as WML in MS (Arnold et al., 2022). Acquiring multiple scans and combining them through super-resolution (SR) could potentially address the limitations of ULF MRI systems, as it can improve image quality beyond the native resolution, allowing for more detailed and sharper visualisation of anatomical structures (Arnold et al., 2022; Van Reeth et al., 2012). SR, an image processing technique to produce high spatial resolution images post-acquisition, works by combining information from multiple lower-resolution images to generate a higher-resolution output, thereby improving image quality and enabling more detailed visualisation of anatomical structures.

A simple method to achieve SR is to use interpolation-based methods. Interpolation-based SR methods increase resolution by estimating voxel values at sub-voxel positions through averaging or interpolating neighbouring voxel values (Van Reeth et al., 2012). In ULF MRI, this approach has been applied

to reslice and average orthogonal acquisitions of simultaneously acquired T2-weighted images and T2 maps (Deoni, O'Muircheartaigh, Ljungberg, Huentelman and Williams, 2022), as well as orthogonal FLAIR acquisitions (Arnold et al., 2022). The latter study employed an interpolation-based SR approximation method, which involved inverting the average diffeomorphism and applying intensity averaging (soft sharpening) (Niaz et al., 2022), referred to here as *ANTs-SR*. However, acquiring multiple planes to generate higher-resolution images significantly extends scan times, reducing clinical practicality. Additionally, because interpolation methods do not create new anatomical information but instead estimate new values from existing neighbouring voxels, they can introduce a blurring effect that reduces image sharpness and may obscure fine details.

Frequency-domain SR approaches aim to increase the resolution of an image by manipulating its frequency components, rather than the voxels in the spatial domain directly (Thapa et al., 2014). In the discrete Fourier domain, motion parameters like planar rotation and shifts can be accurately estimated and then corrected using phase correlation analysis (Vandewalle et al., 2006). Working in the frequency space is valuable as it separates low-frequency components carrying contrast information from high-frequency components containing high-resolution details.

In MRI, spatial information is encoded into the frequency domain during data acquisition and reconstructed into the spatial domain to generate the final image. The relationship between 2D k-space and image space in MRI is described by:

$$\rho(x, y) = \iint S(k_x, k_y) \exp(ik_x x + ik_y y)\, dk_x\, dk_y,$$

where $\rho(x, y)$ is the image in the spatial domain, and $S(k_x, k_y)$ represents the signal in the frequency domain (k-space). The signal, a complex-valued function, describes the amplitude and phase of the signal at each point in k-space. The expression $\exp(ik_x x + ik_y y)$ encodes the spatial information via magnetic field gradients during acquisition. The inverse 2D Fourier transform integrates over all spatial frequencies to recover the image. By acquiring two orthogonal scans and partially sampling the frequency space before performing Fourier transform, it is possible to effectively increase spatial resolution. Alternatively, data can be manipulated during acquisition, reducing the need for full scans as required in interpolation-based approaches. Partially acquiring

frequency-space minimizes complete data acquisition, helping to reduce scan times.

The concept behind combining high-frequency and low-frequency data is shown in Figure 4.14, using a 3T phantom scan as an example. Retaining only the low-frequency components results in an image that maintains contrast but loses spatial resolution, leading to blurring (Figure 4.14A). Conversely, sampling high frequencies preserves only the fine spatial detail, effectively isolating edge information (Figure 4.14B). In anisotropic 3D acquisitions, such as those achievable at ULF, where resolution is lower in the slice direction compared to the in-plane direction, the frequency-space extent along the slice direction is restricted. Instead of acquiring a complete second acquisition, high-frequency data from the outer regions of k-space can be incorporated, improving the overall image quality (Figure 4.14).

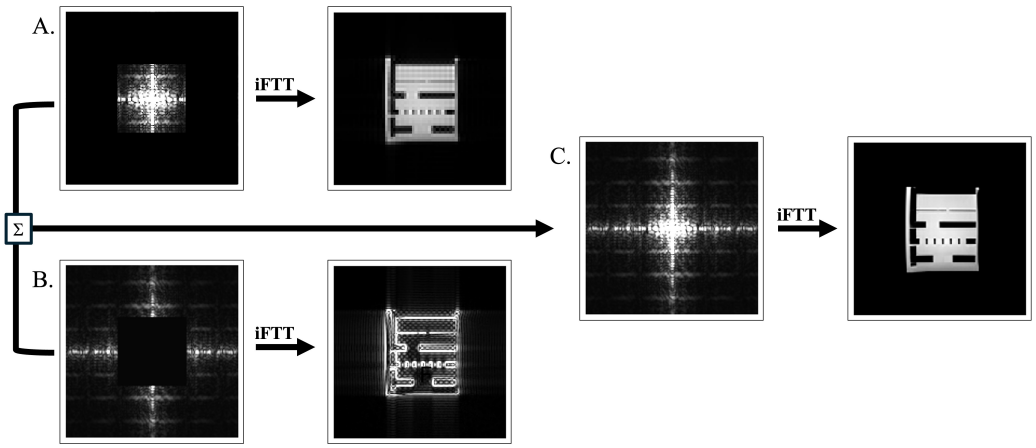


Figure 4.14: An example of how sampling MRI in the frequency domain can affect image quality. A standard hyperfine phantom was scanned at 3T and then down-sampled to 1.6mm isotropic. The image was fast Fourier transformed (FFT), and the resulting frequencies were split. The inverse FFT (IFFT) of A) only the central portion of the frequency space produces a blurred image rich in contrasts. B) The IFFT of the periphery of the frequency space produces images with sharp edges, but the image lacks contrast information between phantom fluid and background. C) The IFFT of the combination of the center and periphery frequencies produces original high-quality image with contrast and sharp edges. This figure is reused from (Donnay et al., 2024).

In this section, I introduce a Fourier-based SR algorithm (FouSR) that leverages high-frequency information from one ULF orthogonal scan direction and combines it with the low-frequency components from another. I hypothesized that by partially sampling data from orthogonal low-resolution scans, this approach would fuse crucial edge and contrast information to improve the delineation of MS lesions and other important features. The work described here has been published previously (Donnay et al., 2024).

4.3.2 Methods

This analysis used same-day paired MRI data from 64 mT and 3 T scanners acquired in a standard Hyperfine phantom and in ten participants with MS, acquired under a natural history protocol at the NIH (Appendix A.1). Participant characteristics and MRI acquisition details are provided in Appendix A.4.3. ULF imaging included two anisotropic FLAIR scans in axial and coronal planes (1.7 mm × 1.7 mm in-plane, 5 mm slice thickness), while 3 T imaging included isotropic FLAIR and T1-weighted scans at 1 mm resolution (Figure 4.15A).

4.3.2.1 Overview of FouSR Approach

All data were resampled to a resolution of 1.7 mm isotropic and zero-padded to create uniform dimensions between HF and ULF (Figure 4.15B). All in-vivo data were skull-stripped using FSL's *BET* for HF scans (Jenkinson et al., 2012) and *SynthStrip* for ULF scans (Hoopes et al., 2022). Then, using the multivariate template construction tool from the Advanced normalisation Tools (ANTs) neuroimaging analysis software, brain-extracted ULF images were used to compute non-linear registration deformation fields to the down-sampled HF images (Avants et al., 2011) (Figure 4.15C). The transformation matrices were then applied to the non-brain extracted images. Resampling the anisotropic images to a common isotropic grid and applying non-linear registration is necessary to ensure accurate spatial alignment, especially since different acquisition orientations can result in different spatial distortions. Spatial misalignments would cause incoherent mixing of frequency information during image combination, degrading the quality of the final image. Each image transformation was performed using nearest-neighbour interpolation and then intensity-normalized to the 98th percentile. This normalisation preserved the original voxel intensities while reducing the influence of extreme intensity outliers. The preprocessed ULF images, acquired in the coronal and axial planes, are referred to as *ULF coronal* and *ULF axial*, respectively.

Figure 4.15D illustrates the image combination process for FouSR. Per the terms of our collaboration agreement with Hyperfine, I did not have access to the raw data, so I used the fast Fourier Transform (FFT) to obtain the data in frequency space. The FFT was applied to the preprocessed ULF coronal and ULF axial scans. In the frequency domain, the inner 64 × 64 matrix

of the ULF coronal was replaced with that from the ULF axial scan. The decision to replace a 64-size matrix was made empirically (Appendix D.3.1). The modified frequency components were then inverse transformed using the iFFT to generate the final image.

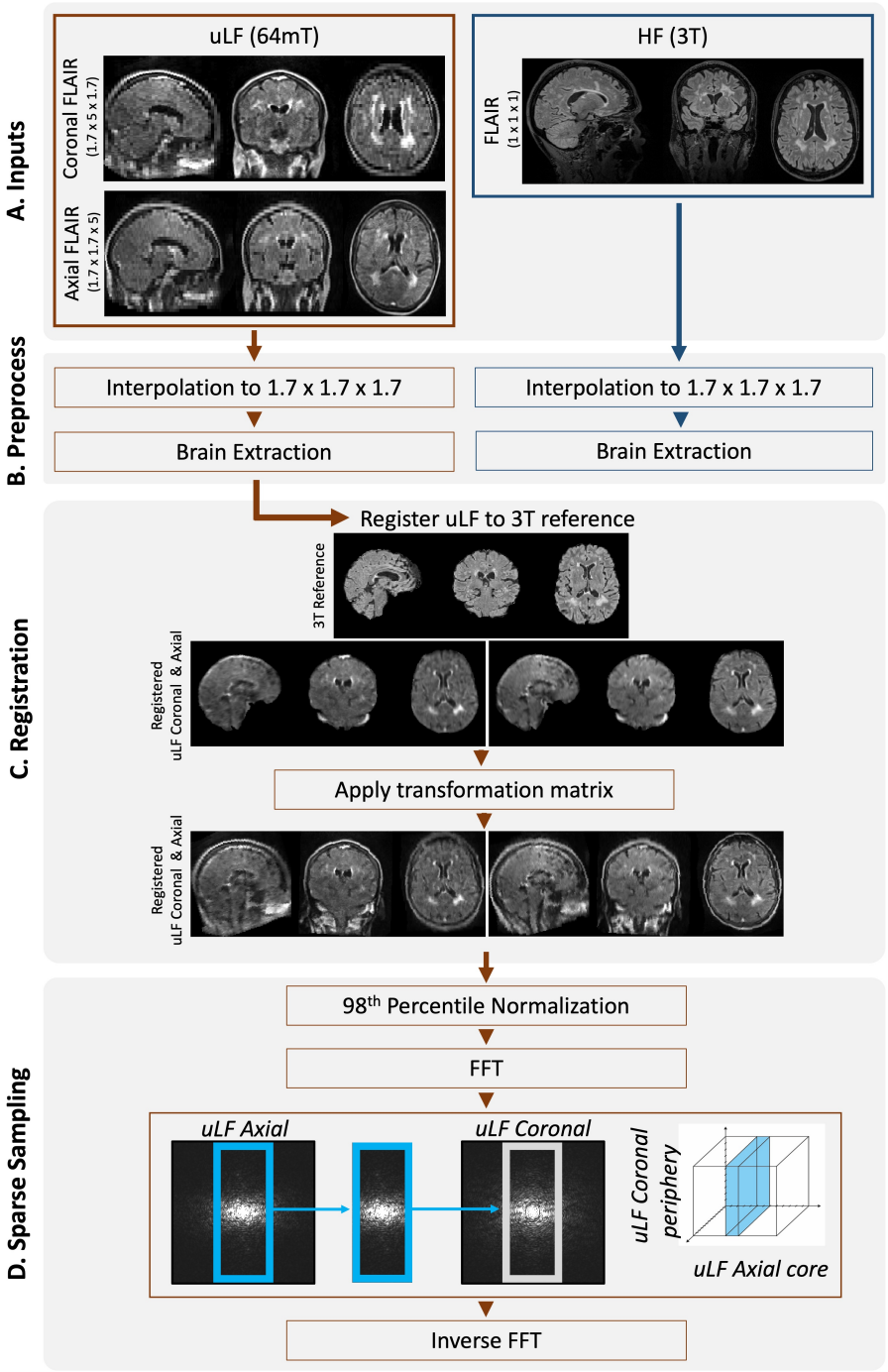


Figure 4.15: Overview of image pre-processing and image combination for Fourier-Based Super Resolution (FouSR). A) ULF (64mT) and HF (3T) MRI were acquired on the same day. ULF FLAIR images were acquired with high resolution in either coronal or axial plane, shown here in tri-planar view. B) All FLAIR images were interpolated to 1.7-mm isotropic and skull-stripped. C) The non-linear transformation from ULF FLAIR scans to the down-sampled HF FLAIR was estimated, and the transformation warp was applied to the non-brain-extracted but up-sampled ULF images. D) The ULF images were Fourier-transformed, and the high-frequency component in the under-sampled direction of ULF Axial FLAIR was then replaced with that from coronal ULF FLAIR. The new frequencies were inversely fast Fourier transformed to image space. This figure is reused from (Donnay et al., 2024).

4.3.2.2 Overview of Comparison Super Resolution Approaches

FouSR images were compared to two interpolation-based SR approximations: *ULF Average* and *ANTs-SR*. The ULF Average was created by performing a voxel-wise average of the ULF coronal and ULF axial scans. This method served as a benchmark for comparing FouSR in both phantom and in-vivo analyses. ANTs-SR was included in the in-vivo analyses. To generate ANTs-SR images, non-skull stripped in-vivo data were resampled to 1.7 mm isotropic and then processed using the default parameters for ANTs multivariate template construction tool, with the HF FLAIR image serving as the reference volume for registration. Laplacian filtering was applied at each iteration, and final images were normalized to their 98^{th} percentile.

4.3.2.3 Quantitative Evaluation of Phantom Data

Using phantom data, SNR and intensity plots were first compared across the ULF Coronal, ULF Axial, ULF Average, and FouSR images. For the SNR calculation, regions of interest (ROIs) were manually annotated in areas of high signal and background, ensuring that regions with artifacts were avoided. The SNR was calculated as the ratio of the mean signal intensity to the standard deviation of the noise. Intensity plots were generated and compared across the three planes for each method.

4.3.2.4 Quantitative Evaluation of In-Vivo Data

On in-vivo data, brain masks were obtained from the down-sampled HF FLAIR (Figure 4.15B) and multiplied by the ULF Coronal, ULF Axial, ULF Average, FouSR, and ANTs SR images. Manual WML masks were annotated on the down-sampled HF FLAIR images and used to calculate the total WML volume. The total WML mask was spatially split into individual clusters for individual lesion analysis. White matter masks were created from HF T1-weighted images using FSL's FAST tool (Zhang et al., 2001) and then linearly registered to the down-sampled HF FLAIR. Total WML masks were subtracted from the white matter masks to create an NAWM ROI. An ROI was manually drawn for each participant in ITKsnap version 3.8.0 (Yushkevich et al., 2006) to measure background noise.

For quantitative evaluation of image quality, the SNR, CNR, and variance of the Laplacian were calculated, similarly to previously described (Arnold et al.,

2022), across SR methods and preprocessed ULF Axial and ULF Coronal. SNR was calculated as described previously (Section 4.3.2.3). CNR was calculated as the ratio of the absolute difference between the mean total lesion and the mean NAWM intensities relative to the standard deviation of background noise. Image sharpness of skull-stripped data was quantified using the variance of the Laplacian, computed by applying a 3D Laplace filter from the *scipy* library in Python (Yushkevich et al., 2006) and then extracting the standard deviation as a measure of variance. Additionally, lesion boundary sharpness was assessed by dilating individual lesion masks and calculating Laplacian variance for these ROI (Lesion Image Sharpness). Lesion conspicuity was calculated as the ratio of the difference between the mean lesion and the mean WM intensity and their sum.

4.3.2.5 Qualitative Evaluation of In-Vivo Data

Three neurologists (S.V.O., M.I.G. and C.T.) with expertise in MS and neuroimaging performed qualitative evaluations of the images, blinded to the image generation method. For each image, the neurologists rated the visibility of WML and the accuracy of sulcal-gyral delineation for four key anatomical landmarks: the central sulcus, parieto-occipital sulcus, calcarine sulcus, and sylvian fissure using a rating scale defined in Table 4.7. To ensure accurate identification of anatomical features and lesions, the down-sampled 3T FLAIR images were provided as a reference during the evaluation process.

Table 4.7: Description of qualitative rating scales for WML and Sulci and Gyri

Score	**Rating**	**WML Description**	**Sulci & Gyri Description***
1	Poor	WML is not visible.	Delineation is not possible. (0-3 pts)
2	Acceptable	WML is visible but hard to identify.	Delineation is possible in some regions. (4-6 pts)
3	Good	WML is visible and easily identifiable.	Delineation is fine in general with PVE related hyperintensities (7-9 pts)
4	Superior	WML-NAWM contrast closer to the 3T.	Sulci-Gyri contrast is sharp with a closer-to-3T delineation (> 9 pts)

Abbreviations: WML = white matter lesion; NAWM = normal appearing white matter
* Rating based on anatomical landmarks (Central sulci, Sylvian fissures, Parieto-occipital sulci, Calcarine sulci)

4.3.2.6 Statistical Analysis

Statistical analyses were conducted using the GraphPad Prism version 9.3.1, GraphPad software, and R version 3.6.3. Before analysis, adherence to the normality assumption was tested using the Shapiro-Wilk test. Repeated-measure ANOVA was used to compare image quality metrics (SNR, CNR, image sharpness, qualitative ratings) across preprocessed ULF images and SR methods. In cases of non-normally distributed data, post hoc multiple comparisons were performed with Dunn's method, and the q ratio was reported. For normally distributed data, Tukey's post hoc multiple comparison test was employed, and the Z statistic was reported. An established significance threshold of a corrected p-value of 0.05 or lower was used to identify statistically significant outcomes. Linear mixed-effects models nested for participants were used to compare image sharpness for individual lesions across different images (*LME4* package in R). Pairwise comparisons were computed using R's *Emmeans* package. Unless stated otherwise, all reported data are presented as mean ± standard error.

4.3.3 Results

4.3.3.1 Quantitative Evaluation: Phantom Data

The FouSR algorithm combined information from both ULF acquisition, retaining comparable image quality of the in-plane dimensions of ULF Coronal and ULF Axial (Figure 4.16). Both ULF Average and FouSR images exhibited comparable signal intensity in areas of high signal to the in-plane ULF scan (Figure 4.16A, B). However, FouSR had a lower SNR (377) compared to ULF Average (614), ULF Coronal (389), and ULF Axial (494). The intensity plots show that FouSR recovered information from ULF Axial, which is visible as sharper slopes (Figure 4.17A) while effectively preserving the signal intensity of the ULF Coronal (Figure 4.17B). Overall, the FouSR retains crucial details from both ULF Axial and Coronal without losing sharp delineations that could otherwise be blurred by averaging the two slices.

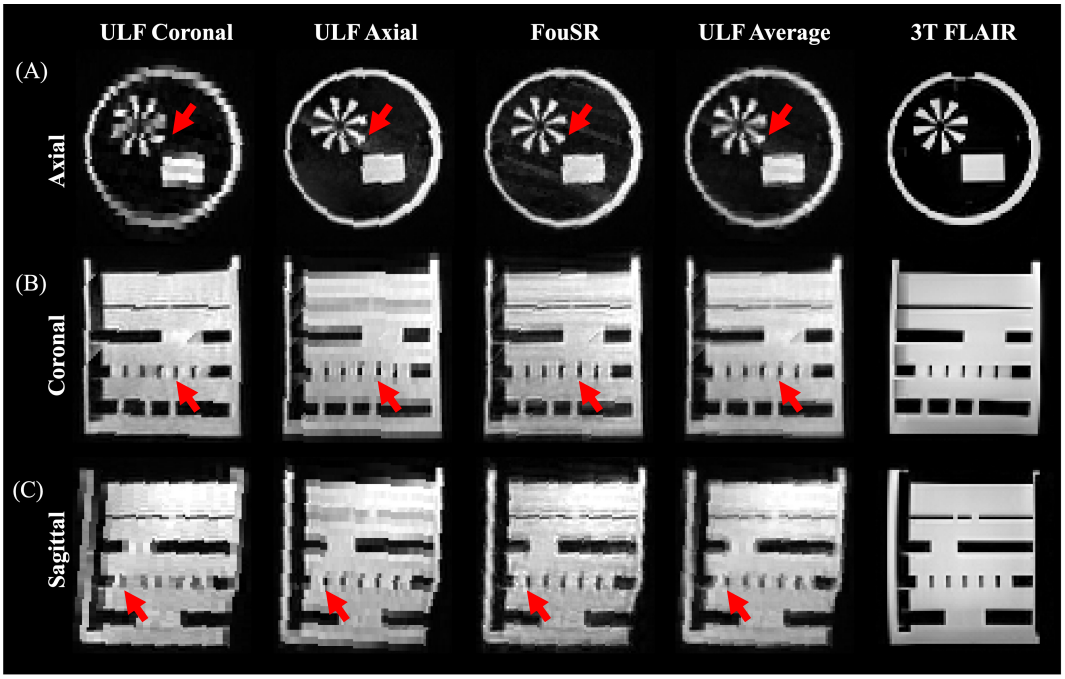


Figure 4.16: Examples of image quality across different methods in a standard Hyperfine phantom. Representative slices are shown in A) axial, B) coronal, and C) sagittal planes for ULF Coronal FLAIR, ULF Axial FLAIR, FouSR, ULF Average, and 3T FLAIR (left to right). Red arrows highlight areas in each plane where FouSR exhibits improved resolution and image sharpness compared to the other methods. This figure is reused from (Donnay et al., 2024).

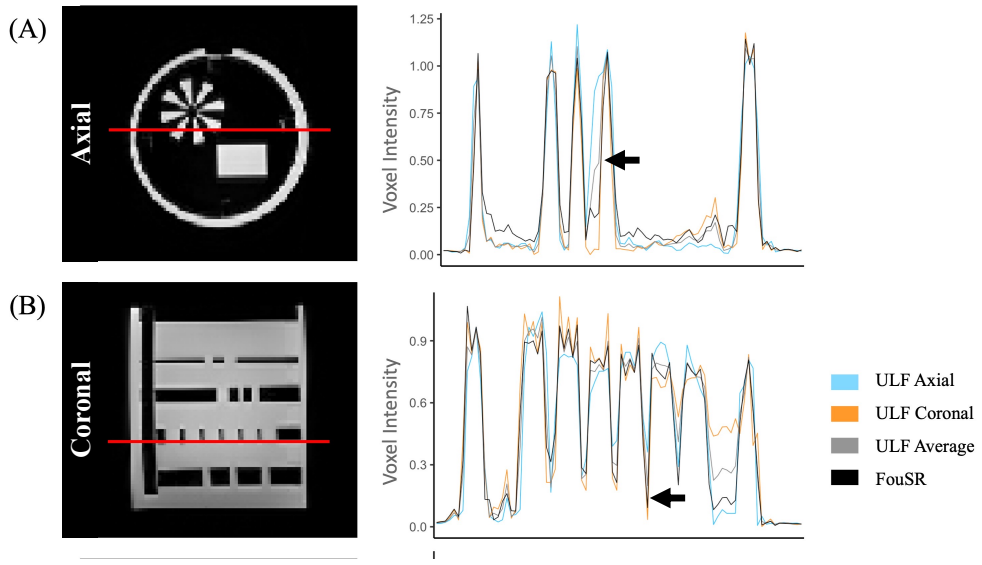


Figure 4.17: Intensity plots of ULF Axial, ULF Coronal, ULF Average, and FouSR. Intensity plots capture voxel intensities of ULF Coronal (orange), ULF Axial (blue), FouSR (black), and ULF Average (gray) along the red line on the 3T FLAIR image on representative slices of a standard Hyperfine phantom. A) axial and B) coronal slices of a standard Hyperfine phantom. The black arrows show examples on each plane where FouSR intensities (black) resemble in-plane line (black and orange overlapping in panel B, black and blue overlapping in panel A) suggesting that information has been recovered in the FouSR algorithm. This figure is reused from (Donnay et al., 2024).

4.3.3.2 Quantitative Evaluation: In-Vivo Data

The participants in this study exhibited varying levels of total lesion burden, with an average of $4313 \pm 3171\ mm^3$. FouSR demonstrated notable improvements in signal definition across all three axes (Figure 4.18). FouSR images closely resembled the high-resolution in-plane orientations of ULF Coronal and ULF Axial scans, which suggests that FouSR was able to recover information from the in-plane voxels of the ULF Coronal and ULF Axial images (Figure 4.18A, columns 1 and 2), resulting in sharper image quality. ULF Average and ANTs SR displayed more prominent blurring, particularly along lesion edges and sulci. However, some artifacts, evident in the ULF Coronal and Axial images (indicated by arrows in Figure 4.18C), were retained in both FouSR and ANTs SR images. These artifacts were less pronounced in the ULF Average images, where the blurring effect was more substantial. Furthermore, FouSR introduced ringing artifacts (Figure 4.18B), which were not present in the other image types, highlighting a potential limitation of this method.

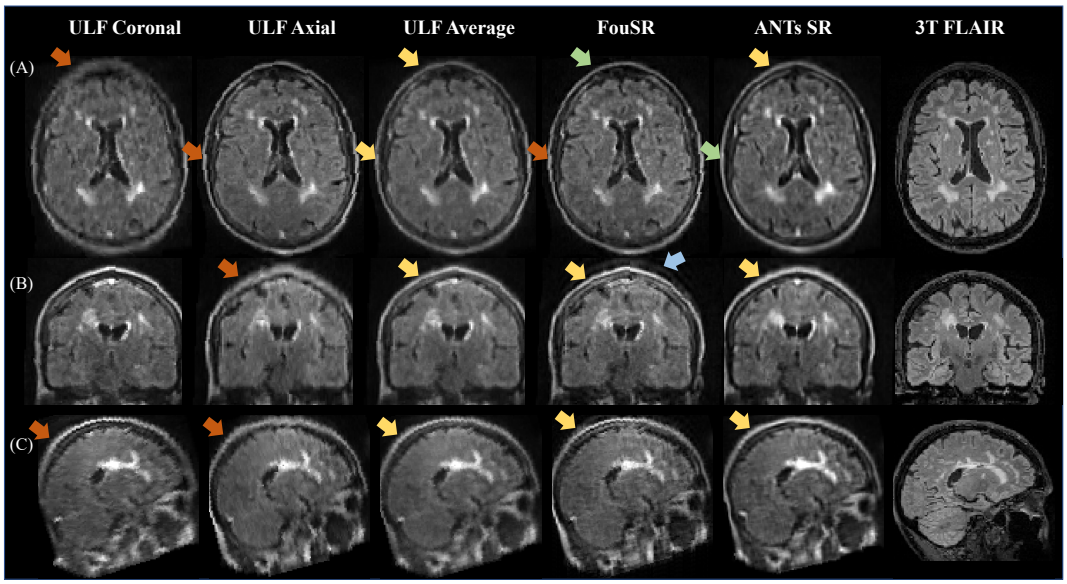


Figure 4.18: Example of different methods applied to a participant with high lesion load. The methods are compared across A) axial, B) coronal, and C) sagittal planes. FouSR demonstrates sharper lesion delineation but also exhibits artifacts indicated by the arrows. The orange arrows highlight partial volume artifacts along the slice direction primarily in ULF Coronal and ULF Axial, which are overall reduced in FouSR, ULF Average, and ANTs SR (yellow arrows) and absent (green arrows) in some areas of ANTs SR and FouSR. The blue arrow indicates a ringing artifact in FouSR. This figure is reused from (Donnay et al., 2024).

Figure 4.19 illustrates the comparison of quantitative metrics across methods. ULF Axial and ULF Average had the highest SNR (54.9 ± 4.8 and 53.2 ± 5.2) of all methods, but only ULF Axial's SNR was significantly higher

than FouSR (48.0 ± 4.0, mean difference = 6.9, $p = 0.005$). ULF Coronal exhibited the lowest SNR (36.7 ± 3.8), which was significantly lower than ULF Axial (mean difference = -18.2, $p = 0.01$), ULF Average (mean difference = -16.5, $p < 0.001$), and FouSR (mean difference = -11.3, $p = 0.037$) (Figure 4.19A). ULF Average had the highest mean CNR (12.5 ± 2.8), which was only significantly higher than ULF Coronal (9.2 ± 1.9, $z = 3.1$, $p = 0.019$) (Figure 4.19B). Lesion conspicuity was higher in FouSR (0.12 ± 0.012) compared to ULF Axial (0.11 ± 0.018, mean difference = -0.007, $p = 0.002$) but not significantly different from other methods.

FouSR demonstrated greater sharpness relative to all other image types (Figure 4.19C). Specifically, FouSR showed an 18% and 23% improvement in image sharpness compared to ULF Average and ANTs SR, respectively. To assess lesion sharpness, a linear mixed model was applied, nested for participants with individual lesions. To meet normality assumptions, lesion sharpness values were log-transformed. FouSR exhibited significantly higher log lesion sharpness (-0.971 ± 0.026) than ULF Average (-1.37 ± 0.03, $t(543) = 10.2$, $p < 0.0001$). ULF Average exhibited lower lesion sharpness compared to both ULF Axial (-1.14 ± 0.03, $t(543) = -7.51$, $p < 0.0001$) and ULF Coronal images (-1.05 ± 0.03, $t(543) = -8.83$, $p < 0.0001$), likely due to boundary-blurring effects from averaging (Figure 4.19D).

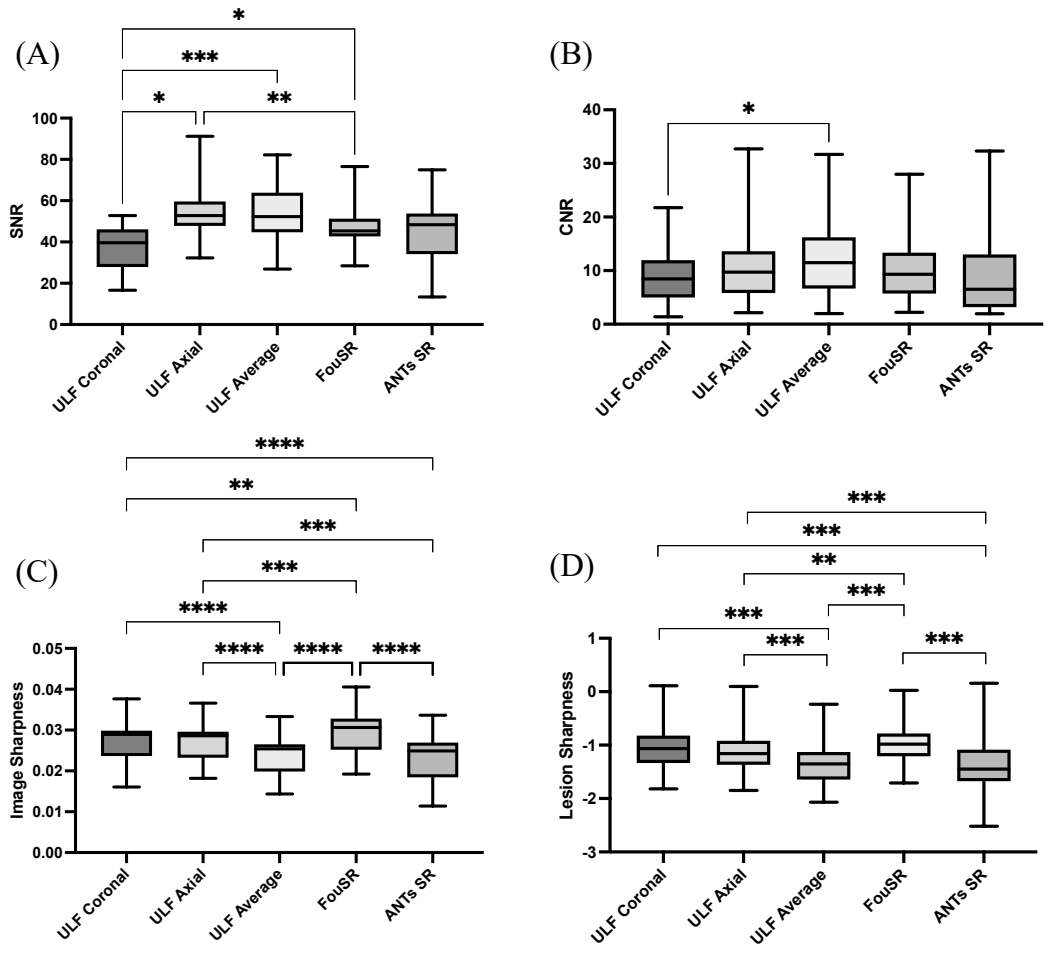


Figure 4.19: Comparison of quantitative scan quality metrics in in-vivo data ($n = 10$) across methods. A) SNR, B) CNR, C) Image sharpness, and D) Lesion sharpness is compared across methods. Only significant pairwise differences are presented. This figure is reused from (Donnay et al., 2024).

4.3.3.3 Qualitative Evaluation: In-Vivo Data

Figure 4.20 shows the results from qualitative neurologist ratings. WML visibility was similar across all methods. While ANTs SR and FouSR had the highest mean WML ratings (2.68 ± 0.24 and 2.68 ± 0.26), the only significant difference was observed between ANTs SR and ULF Coronal (2.3 ± 0.24, mean difference $= 0.39$, $p = 0.004$) (Figure 4.20A). FouSR did not have statistically different sulci and gyri ratings compared to any other method (Figure 4.20B). ANTs SR (2.43 ± 0.16), ULF Average (2.2 ± 0.10), and ULF Axial (2.2 ± 0.12) all had significantly higher ratings compared to ULF Coronal (1.65 ± 0.09) (Figure 4.20B).

While performing ratings, the neurologists observed that ULF Average and ANTs SR exhibited more uniform intensities across different tissue types, par-

ticularly in the white matter. This uniformity is useful for accurately identifying anatomical landmarks and lesions.

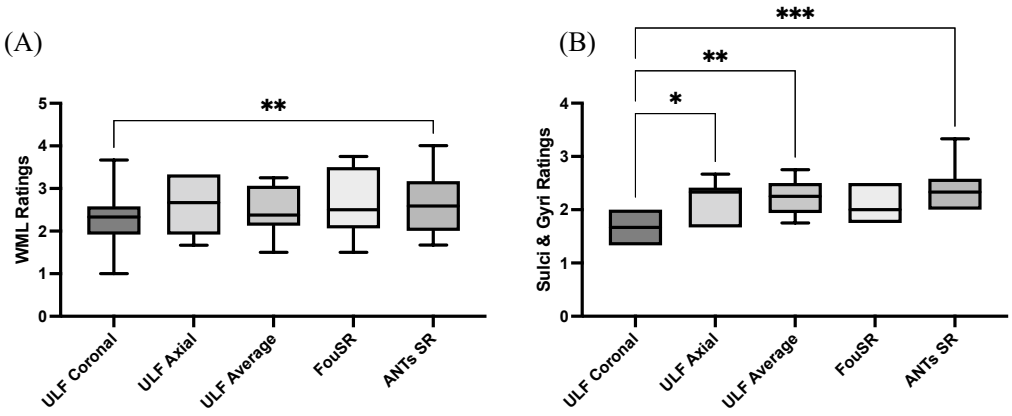


Figure 4.20: Comparison of qualitative ratings for different methods. Average ratings across all raters for A) WML and B) Sulci and Gyri. Only significant pairwise differences are displayed. This figure is reused from (Donnay et al., 2024).

4.3.4 Discussion

In this section, I introduce FouSR, a super-resolution (SR) algorithm developed to improve the resolution of ULF MRI scans by leveraging high-frequency information from an additional orthogonal scan direction. Images reconstructed using FouSR demonstrated superior boundary sharpness compared to ULF coronal, ULF axial, and ULF Average while maintaining similar quantitative image metrics, such as SNR and CNR.

FouSR was compared with two other image combination methods: ULF Average and ANTs SR. All three methods use the ANTs multivariate template construction tool to some extent (Avants et al., 2011). For FouSR and ULF Average, the tool is only used to obtain an optimized registration matrix, chosen for its superior registration compared to other open-source tools that were tested. ANTs SR, selected based on its successful application in previous ULF SR studies (Deoni, O'Muircheartaigh, Ljungberg, Huentelman and Williams, 2022), uses the tool's default output.

The main limitations observed in the images generated by FouSR were inhomogeneous texture within tissues and the presence of ringing artifacts. These issues may arise from several factors. For instance, using nearest-neighbour interpolation may introduce heterogeneity. While this was chosen to preserve SNR and image sharpness, the resulting inhomogeneity can make it more chal-

lenging to distinguish noise from subtle anatomical features, such as small lesions.

While FouSR demonstrated good delineation of MS lesions, as evidenced by high neurologist WML ratings and sharper lesion boundaries, the blurring observed in ULF Average and ANTs SR was favourable to the neurologists. This is because blurring resulted in more uniform intensity values within tissue types, which can aid in identifying anatomical structures. ANTs SR, in particular, provided more consistent intensity values across tissue types. The clear demarcation of homogeneous intensity values between white matter and WML is beneficial for lesion detection, despite the trade-off in sharpness. However, ANTs SR's incorporation of a sharpening filter improved the delineation of lesion edges compared to ULF Average.

FouSR also exhibited a higher occurrence of Gibbs ringing artifacts. These artifacts can be attributed to differences in the phase and slice encoding directions between the ULF coronal and ULF axial images at the time of acquisition, resulting in discrepancies in off-resonance components, distortions, and B_0 field inhomogeneities. Given the naïve combination of ULF coronal and axial images (i.e., simply replacing frequency lines), FouSR is highly sensitive to spatial and intensity inconsistencies, with any mismatches potentially resulting in undesired artifacts.

The sharp truncation in the frequency components of FouSR makes it more sensitive to registration errors compared to simple averaging. Even small registration errors due to distortion and motion can significantly disrupt the alignment between high- and low-frequency components, whereas averaging smooths out these misalignments. In-vivo ULF acquisitions exhibited more noticeable distortions, especially in the occipital region, compared to high-field scans. Distortions in coronal and axial acquisitions were dissimilar, largely due to swapping of the phase and slice encoding directions. In contrast, ULF phantom acquisitions showed minimal distortion artifacts. However, distortions in in-vivo scans occasionally affected the quality of non-linear registrations, and residual mismatches led to ringing artifacts. These artifacts were less noticeable in the ULF Average images, likely due to the smoothing effect of averaging, though at the cost of boundary sharpness.

FouSR's image combination process was performed in the frequency space after zero-padding, interpolation to isotropic resolution, and registration to a paired HF scan. However, spatially transforming an image before Fourier

transformation can diminish image quality, as interpolation modifies the spatial distribution of voxels, altering the frequency content. Methods like zero-padding and nearest-neighbour interpolation may introduce abrupt transitions between voxels, leading to unwanted artifacts within the Fourier domain. This manifested in the FouSR outputs as pronounced Gibbs ringing and noticeable granularity.

Despite these limitations, one significant advantage of FouSR over ANTs SR and ULF Average is its potential for easier implementation directly on the scanner. While the FouSR algorithm has been described here as a post-acquisition image processing technique, it can also be integrated into the acquisition process. In this case, the additional orthogonal scan would only need to capture the high-frequency components, offering a potential $\sim 3\times$ improvement in slice resolution. This approach would result in an estimated $\sim 60\%$ increase in scan time compared to a single-plane acquisition, which is still 20% shorter than the time required for two full orthogonal scans. Furthermore, the overall data size would increase by only 40% relative to a single acquisition and decrease by 33% compared to two full acquisitions.

Many of the described limitations result from the implementation of FouSR by reverse-engineering frequency space from already-processed magnitude images. This represents a fundamental limitation of this work: due to proprietary scanner restrictions and our agreement with Hyperfine, we did not have access to raw complex k-space data, including phase information. This necessitated post-acquisition interpolation and inverse Fourier transformation, which introduced artefacts as discussed above. Phase information is essential for accurate motion estimation and correction in the frequency domain, as well as for precise image registration.

Thus, many current limitations of FouSR could be addressed if the algorithm were implemented directly on the scanner. Implementing FouSR on the Hyperfine System could be achieved by modifying the k-space trajectory. However, due to the proprietary nature of the scanner interface, this could only be achieved through collaboration with the manufacturer, which was not possible during this study. Incorporating B_0-maps and navigator-based motion correction could further improve the SR outputs, particularly if integrated into on-scanner reconstruction (Liu, Beck, Filippini, van Gelderen, de Zwart, Norato, Sati, Al-Louzi, Kolb, Donadieu, Morrison, Duyn and Reich, 2021). While B_0-maps can be acquired post-acquisition, acquiring and reconstructing FouSR

images during data acquisition with real-time motion correction would address many current challenges and significantly enhance SR quality.

Incorporating FouSR during acquisition would not only address issues related to post-processing interpolation but also allow inclusion of navigator-based corrections in k-space prior to SR combination. Navigators are short sequence elements incorporated into the image sequence within the echo time and offer real-time measurements of motion and distortion (Duan et al., 2015; Ehman and Felmlee, 1989; Liu et al., 2020; Maclaren et al., 2013; Welch et al., 2004). This real-time processing allows for immediate adjustments to the imaging system, reducing artifacts caused by dynamic perturbations and enhancing overall image quality (Liu et al., 2020; van der Kouwe et al., 2006).

This study compared simple interpolation and Fourier-based SR algorithms, although many previous studies have investigated DL to improve ULF MRI resolution (Ayde et al., 2022; de Leeuw den Bouter et al., 2022; Iglesias et al., 2023a; Lau et al., 2023). Learning-based SR techniques use CNNs to detect complex patterns, enabling high-resolution image generation from low-resolution inputs. DL-based SR can also reduce noise and mitigate Gibbs ringing (Lau et al., 2023) and simulate HF contrast (Iglesias et al., 2023a). However, these methods face challenges, such as the lack of large paired HF and ULF datasets. Down-sampling and noise addition are often used to simulate ULF conditions but struggle to capture unique noise, contrast, and spatial distortions. FouSR does not depend on DL or synthetic data; it enhances resolution of existing ULF acquisitions, making it adaptable to various contrasts and disease types. Changes are limited to frequency space sampling strategies if acquisition resolution is modified.

There are some limitations related to the design of this study. Only 10 in-vivo scans were analysed, limiting generalizability. WML burden varied within the cohort and could influence quantitative and qualitative metrics; however, this was not explored due to limited sample size. Inaccuracies from downsampling and registration could impact SNR, CNR, and lesion conspicuity, particularly for smaller lesions. Additionally, standard deviation of noise used in CNR and SNR calculations was derived from ROIs drawn on scan background. Variability in noise measures can arise from artifacts in phase and slice directions, which differ across images. ROIs were carefully placed to avoid regions without noise due to compressed-SENSE acquisition. Alternative sampling

schemes after registration or combination of different orthogonal acquisitions (e.g., sagittal) were not explored.

In this section, the potential of FouSR in enhancing ULF scan resolution was demonstrated. In a small dataset, FouSR improved the delineation of MS lesions compared to single-plane acquisitions, potentially enhancing ULF MRI's ability to detect and monitor subtle lesional changes. Implementing FouSR directly on the scanner could further improve image quality and clinical utility without the need for acquiring multiple full images in orthogonal planes.

Chapter 5

Preliminary Application of Ultra-Low Field MRI for Tracking Multiple Sclerosis White Matter Lesions

Contents

5.1 Introduction 217
5.2 Methods 218
5.3 Results 221
5.4 Discussion 228

5.1 Introduction

In multiple sclerosis (MS), white matter lesion (WML) volume is a commonly reported measure in research and routine care, serving as an objective surrogate marker of disease burden. WML burden is consistently used to evaluate the efficacy of disease-modifying therapies in trials, and cross-sectional measures of WML have been consistently shown to correlate modestly with clinical disability (Fisniku et al., 2008; Rovaris et al., 2003). Longitudinal changes in WML are more sensitive indicators of disease activity and are predictive of future progression. Subtraction MRI at conventional field strengths has demonstrated that lesion activity over one year can predict longer-term disability outcomes up to four years (Liguori et al., 2011). Furthermore, the detection of new or enlarging lesions on serial MRI is used as radiological evidence of dissemination in time (DIT), a key diagnostic criterion for MS. Tracking WML evolution over

time is therefore a well-established component of both clinical monitoring and research in MS (Moraal et al., 2009).

Given the emerging role of ultra-low field (ULF) MRI as an accessible and portable neuroimaging alternative to conventional MRI (cMRI) scanners, establishing whether ULF-derived lesion metrics capture clinically relevant information observed on cMRI is a critical step toward understanding their potential utility in clinical settings. In Chapter 4.3.4, I described a set of tools for WML segmentation and longitudinal lesion tracking at ULF MRI. While these tools showed technical feasibility, it is not yet known whether lesion measures derived from ULF reflect disease severity or activity in ways comparable to cMRI. In this chapter, I present preliminary, exploratory applications of these tools to better understand their potential clinical relevance. First, I assess whether ULF-derived lesion volumes from PLAn segmentations align with radiologically defined lesion burden categories (mild, moderate, or severe) and whether they capture clinical correlations observed with 3T-derived metrics. Second, I examine whether longitudinal changes detected using ULF subtraction imaging are related to clinical disability. Clinical disability scores for these analyses were extracted by a colleague (A.T) at the National Institutes of Health (NIH).

5.2 Methods

5.2.1 Clinical Correlates of Cross-sectional Total Lesion Volume

The dataset used for this analysis is a subset of the validation cohort described in Chapter 4.1, only including participants with a confirmed diagnosis of MS and a clinical evaluation within 6 months of the MRI scan. This cross-sectional cohort comprises participants who underwent same-day 3T and ULF MRI (Appendix A.5). Participants were radiologically assigned into one of three lesion burden categories (1-3 corresponding to mild, moderate, or severe), following the criteria previously defined in Chapter 4.3.4, Section 4.1.2.1. For each MRI scan, clinical disability scores from the closest clinical evaluation to the MRI scan were extracted, where available. These included the Expanded Disability Status Scale (EDSS), Scripps Neurological Rating Scale (Scripps), 9-Hole Peg Test (9HPT), Paced Auditory Serial Addition Test (PASAT), and Symbol Digit Modalities Test (SDMT). As described previously, all participants had

reference 3T and ULF WML segmentations generated using a variety of automatic segmentation algorithms. Full details on the generation of reference and automatic segmentation masks are provided in Chapter 4.3.4, sections 4.1.2.2 and 4.1.2.3, respectively.

To assess whether ULF-derived lesion segmentations captured the same associations observed at 3T, total WML volumes were first tested for their ability to differentiate between radiologically assigned lesion burden categories, and their relationship with clinical disability scores was then examined. Both analyses were initially performed using 3T Reference segmentations to establish a benchmark within this dataset and then repeated using two ULF segmentation masks (ULF Aided and PLAn 2 FLAIR). The ULF Aided masks represent the upper bound of segmentation performance at ULF, as they were manually annotated with reference to 3T scans. PLAn 2 FLAIR was included to evaluate whether an automatic tool could approximate the lesion burden and clinical relevance captured by the 3T Reference or ULF Aided masks and was selected for its consistent performance among the automatic methods tested in Chapter 4.1. Appendix E.1 extends these analyses to include all automatic WML segmentation methods evaluated in Chapter 4.1.

5.2.1.1 Statistical Analysis

Differences in total WML volumes across radiologically defined lesion burden categories (independent groups) were analysed using the Kruskal-Wallis test, reporting the test statistic H. Significant results were followed up with Dunn's post hoc multiple comparisons test, reporting Z statistics and p-values adjusted for multiple comparisons. A two-sided adjusted p-value < 0.05 was considered statistically significant. Spearman correlations and age-adjusted linear regression models were applied to examine relationships between clinical scores and total WML volume.

5.2.2 Clinical Correlates of Longitudinal Lesion Evolution

This analysis used a subset of the cohort from Chapter 4.2, including only participants with confirmed MS and clinical scores within 6 months of their baseline and follow-up MRI. I examined whether lesion changes detected with ULF subtraction imaging were associated with clinical disability changes, using 3T

and ULF radiological assessments from Chapter 4.2. As previously described (Chapter 4.2.2.2), radiological assessments included the total number of new lesions (NL), enlarging lesions (EL), and shrinking lesions (SL) per participant, identified at both 3T and ULF. To reflect the net balance of lesion progression and regression, two additional participant-wise metrics were derived: positive disease activity (PDA) and overall disease activity (DA). PDA was defined as the sum of NL and EL. DA was calculated by subtracting the total number of SL from the sum of total NLs and ELs. These composite measures capture the fact that participants may exhibit both worsening (from NLs and/or ELs) and improvement (from SL) simultaneously, which would not be captured when treating NL, EL, and SL counts separately. Negative disease activity (NDA) was defined as the number of total SL.

Clinical disability scores were extracted from the clinical visit closest to the MRI scan date, within a 6-month window at both baseline and follow-up. After checking normality assumptions, Spearman correlations were calculated between participants' longitudinal changes in clinical scores and lesion activity measures (number of NLs, ELs, SLs, PDA, and DA) derived separately from 3T and ULF MRI.

In addition, changes in clinical disability scores over time were visualized using line plots. Participants were grouped based on whether ULF MRI correctly or incorrectly captured case-level lesion activity compared with the 3T reference (Chapter 4.2). For participants with true positive ULF case-level activity (ULF activity confirmed on 3T), lesion activity was further categorized as:

- Case-level PDA: at least one NL or EL detected on ULF and no SLs.
- Case-level NDA: at least one SL detected on ULF, with no NL or EL.
- Case-level stable: absence of NL, EL or SL.

For participants with false negative ULF case-level activity (lesion activity not detected on ULF but detected on the reference 3T), changes in disability scores were also plotted, though only case-level false negative PDA was observed in this group.

Given that these subgroup sizes were small, descriptive visualization (line plots) were used without statistical testing.

5.3 Results

5.3.1 Clinical Correlates of Cross-sectional Total Lesion Volume

In 72 participants with MS (mean age 49 ± 1.5 years; 52 female), WML volume increased with radiologically defined lesion burden across all segmentation methods (Figure 5.1). Kruskal-Wallis tests showed significant group differences for 3T Reference ($H = 47$, $p < 0.001$), ULF Aided ($H = 47$, $p < 0.001$), and PLAn 2 FLAIR ($H = 42$, $p < 0.001$) volumes. For all methods, post hoc Dunn's tests adjusted for multiple comparisons indicated significant volume increases from burden score 1 to 2 (3T: $Z = 4.8$, $p < 0.001$; ULF Aided: $Z = 4.4$, $p < 0.001$; PLAn 2 FLAIR: $Z = 4.5$, $p < 0.001$), from 1 to 3 (3T: $Z = 6.0$, $p < 0.001$; ULF Aided: $Z = 6.0$, $p < 0.001$; PLAn 2 FLAIR: $Z = 5.8$, $p < 0.001$), but not from 2 to 3 (3T: $Z = 2.0$, $p = 0.12$; ULF Aided: $Z = 2.1$, $p = 0.11$; PLAn 2 FLAIR: $Z = 2.1$, $p = 0.10$). Importantly, most participants fell into lesion burden category 1 ($n = 39$) and lesion burden category 2 ($n = 22$), while only 11 participants had lesion burden category 3.

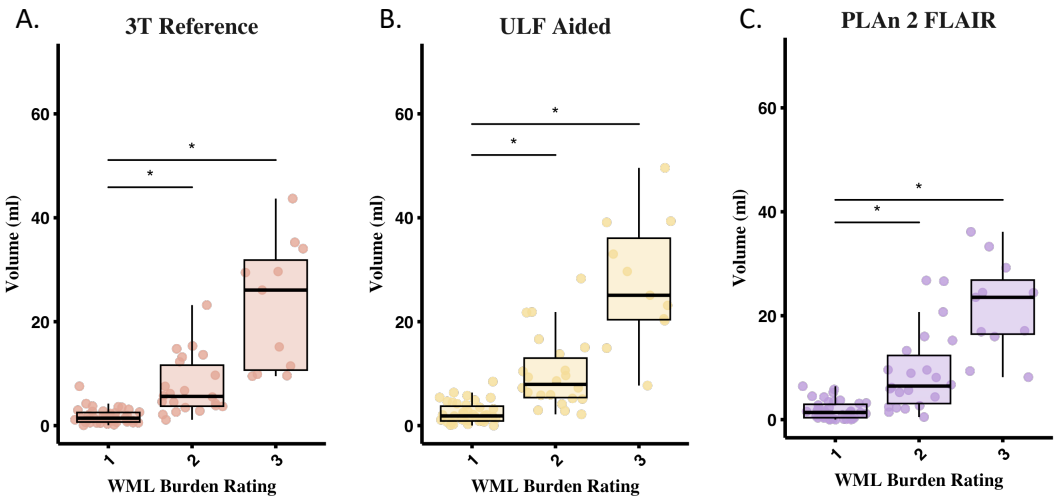


Figure 5.1: Total WML volumes across radiologically defined lesion burden categories. Total WML volumes measured using A) 3T reference segmentations, B) ULF aided segmentations, and C) automatic PLAn 2 FLAIR segmentations, shown across predefined radiological lesion burden categories (1-3)

For all participants, clinical scores were extracted from the closest available clinical visit (mean difference between MRI scan and clinical visit date: 1.34 ± 0.7 days; range: 0-42 days). Table 5.1 summarizes the clinical characteristics of sub-cohorts defined by the availability of each clinical score.

Table 5.1: Sub-cohorts of participants with available clinical scores. Numeric variables are shown as mean ± standard error (SE) with range [min, max] where applicable. Counts are shown for categorical variables.

		Sub-cohorts by clinical score availability				
		EDSS	**Scripps**	**9HPT DH**	**PASAT**	**SDMT**
Participant Characteristics	# of participants	70	70	30	23	27
	Age in years	48.9 ± 1.56	48.9 ± 1.56	47.4 ± 2.07	45.6 ± 2.53	46.7 ± 2.25
	Sex (Female, Male)	50, 20	50, 20	22, 8	16, 7	20, 7
	Sub-Diagnoses (RRMS, SPMS, PPMS)	44, 11, 4	44, 11, 4	21, 3, 2	16, 2, 2	18, 3, 2
	Days between MRI and clinical evaluation	1.3±0.69 [0, 42]	1.3±0.69 [0, 42]	1.3±0.81 [0, 20]	1.1±0.89 [0, 20]	1.4 ± 0.9 [0, 20]
MRI Characteristics	64mT WML volume (mL)	8.6±1.29 [0, 49.6]	8.6±1.29 [0, 49.6]	8.6±2.22 [0.2, 49.6]	5.6±1.51 [0.2, 28.3]	7.9±2.16 [0.2, 49.6]
	3T WML volume (mL)	6.8±1.13 [0.1, 43.7]	6.8±1.13 [0.1, 43.7]	6.7±1.87 [0.5, 43.7]	4.4 ± 1.2 [0.5, 23.2]	5.7±1.51 [0.5, 34.1]

Abbreviations: WML = White Matter Lesion; EDSS = Expanded Disability Status Scale; Scripps = Scripps Neurological Rating Scale; 9HPT DH = 9-Hole Peg Test Dominant Hand; PASAT = Paced Auditory Serial Addition Test; SDMT = Symbol Digit Modalities Test

3T reference WML segmentation volumes demonstrated positive correlations with participant age ($\rho(70) = 0.39$, $p < 0.001$), EDSS ($\rho(68) = 0.33$, $p = 0.006$), and negative correlations with Scripps scores ($\rho(68) = -0.29$, $p = 0.01$) and SDMT ($\rho(25) = -0.43$, $p = 0.02$) (Figure 5.2A). Both ULF-based segmentation volumes captured similar relationships: with age (ULF Aided: $\rho(70) = 0.38$, $p < 0.001$; PLAn 2 FLAIR: $\rho(70) = 0.38$, $p < 0.001$), EDSS (ULF Aided: $\rho(68) = 0.34$, $p = 0.004$; PLAn 2 FLAIR: $\rho(68) = 0.37$, $p = 0.002$), Scripps (ULF Aided: $\rho(68) = -0.32$, $p = 0.008$; PLAn 2 FLAIR: $\rho(68) = -0.33$, $p = 0.005$), and SDMT (ULF Aided: $\rho(25) = -0.46$, $p = 0.01$; PLAn 2 FLAIR: $\rho(25) = -0.41$, $p = 0.03$) (Figure 5.2A).

Similarly, linear regression confirmed a significant association between age and 3T reference lesion volume ($\beta_{\text{Volume}} = 0.38$, $p = 0.02$, Figure 5.2B), and age-adjusted models revealed independent associations between 3T reference lesion volume and both EDSS ($\beta_{\text{Volume}} = 0.07$, $p = 0.002$, Figure 5.2C) and Scripps scores ($\beta_{\text{Volume}} = -0.64$, $p = 0.001$, Figure 5.2D). These relationships were also captured using ULF-derived segmentation volumes. Age was significantly

associated with both ULF Aided ($\beta_{\text{Volume}} = 0.35$, $p = 0.01$) and PLAn 2 FLAIR volumes ($\beta_{\text{Volume}} = 0.41$, $p = 0.02$, Figure 5.2B). In age-adjusted models, ULF Aided and PLAn 2 FLAIR volumes were both significantly associated with EDSS ($\beta_{\text{Volume}} = 0.06$, $p = 0.002$; $\beta_{\text{Volume}} = 0.08$, $p = 0.003$, Figure 5.2C) and Scripps scores ($\beta_{\text{Volume}} = -0.62$, $p < 0.001$; $\beta_{\text{Volume}} = -0.76$, $p < 0.001$, Figure 5.2D). No significant associations were observed between SDMT and lesion volume in any of the age-adjusted models. Appendix E.1 provides full Spearman and linear regression results for all WML segmentation methods evaluated in Chapter 4.1.

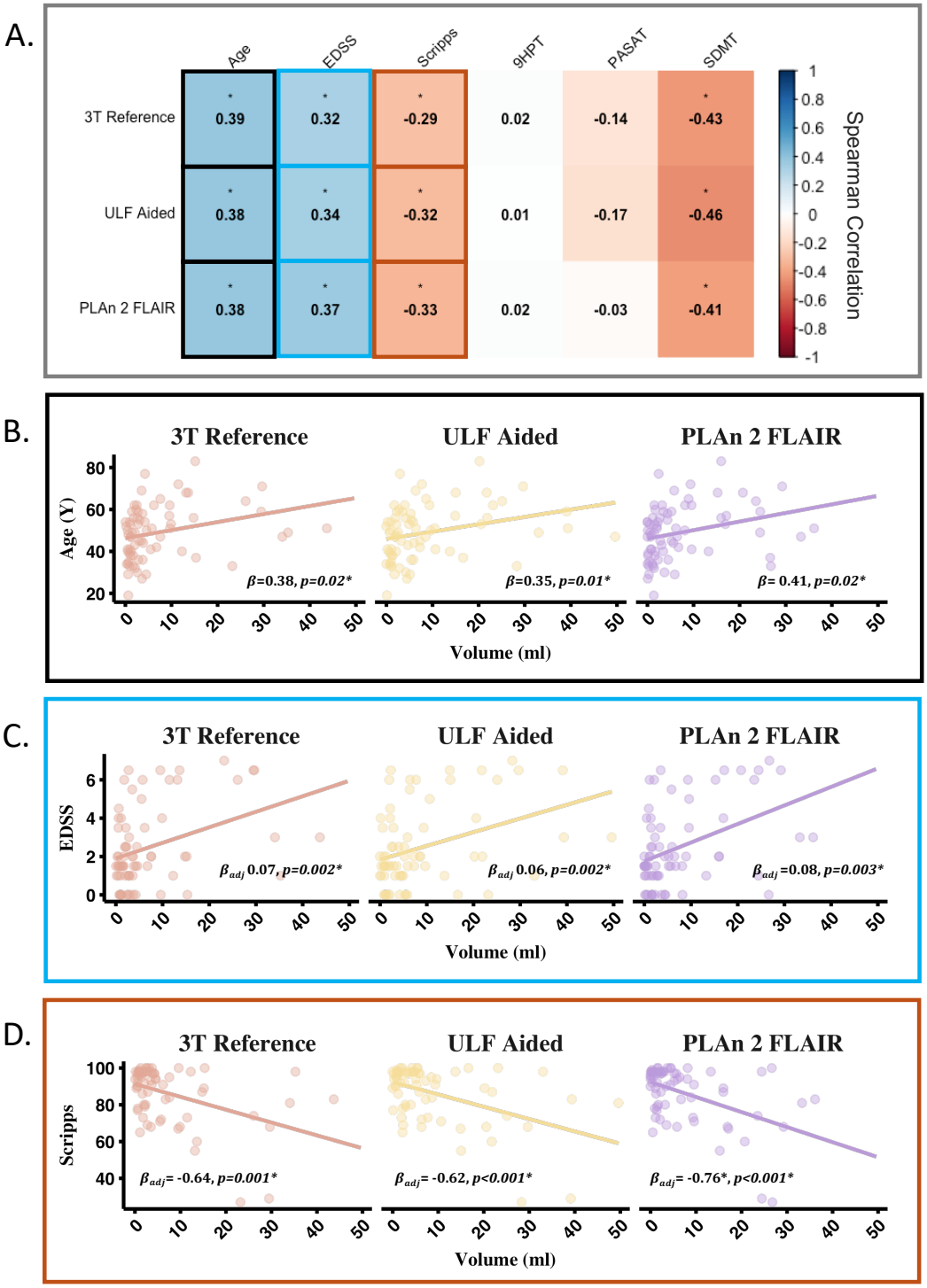


Figure 5.2: Clinical correlations with total WML volume. A) Spearman correlation matrix showing associations between clinical measures (age and clinical scores, shown along the horizontal axis) and total lesion volumes obtained from 3T reference, ULF Aided and PLAn 2 FLAIR segmentation methods. Significant correlations are marked with an asterisk (*). B) Linear regression models of age versus total lesion volume for reference segmentation masks at 3T, ULF manual masks, and automatic PLAn 2 FLAIR segmentations. C) Age-adjusted linear regression models of EDSS as a function of total lesion volume measured by 3T reference masks, ULF manual masks, and PLAn 2 FLAIR masks. D) Age-adjusted linear regression models of Scripps scores as a function of total lesion volume measured by 3T reference masks, ULF manual masks, and PLAn 2 FLAIR masks.

5.3.2 Clinical Correlates of Longitudinal Lesion Evolution

In the longitudinal subset of 21 participants with MS (mean age 50.6 ± 2.0 years; 71% female), EDSS and Scripps clinical scores were available for most participants ($n = 20$, $n = 19$, respectively) at both baseline and follow-up. Fewer participants had PASAT, 9HPT, and SDMT scores at both time points ($n < 10$). Table 5.2 summarizes the clinical characteristics of sub-cohorts defined by score availability.

Table 5.2: Clinical characteristics of MS participants with longitudinal radiological lesion activity assessments, grouped by availability of longitudinal clinical scores. Numeric variables are shown as mean ± standard error (SE) with range [min, max] where applicable. Counts are shown for categorical variables.

		Sub-cohorts by clinical score availability				
		EDSS	**Scripps**	**9HPT DH**	**PASAT**	**SDMT**
Participant Characteristics	# of participants	20	19	9	8	8
	Baseline Age in years	51.2 ± 2	51.6 ± 2.1	51.6 ± 3.6	51.8 ± 4	51.8 ± 4
	Sex (Female, Male)	17,1,2	16,1,2	8,0,1	7,0,1	7,0,1
	Sub-Diagnoses (RRMS, SPMS, PPMS)	14	13	8	7	7
Baseline MRI – Clinical Visit Interval (days)		16.1 ± 9.9 [0, 168]	16.6 ± 10.4 [0, 168]	23.6 ± 18.5 [0, 168]	22 ± 20.9 [0, 168]	22 ± 20.9 [0, 168]
Follow-up MRI – Clinical Visit Interval (days)		7.3 ± 6.1 [0, 121]	7.7 ± 6.4 [0, 121]	2.3 ± 2.2 [0, 20]	2.6 ± 2.5 [0, 20]	2.6 ± 2.5 [0, 20]

Abbreviations: WML = White Matter Lesion; EDSS = Expanded Disability Status Scale; Scripps = Scripps Neurological Rating Scale; 9HPT DH = 9-Hole Peg Test Dominant Hand; PASAT = Paced Auditory Serial Addition Test; SDMT = Symbol Digit Modalities Test

At both field strengths, radiologically detected lesion activity was significantly associated with a change in the Scripps score (Figure 5.3). DA was negatively associated with change in Scripps (3T: $\rho(18) = -0.46$, $p = 0.04$; ULF: $\rho(18) = -0.62$, $p = 0.004$) while the number of SL was positively associated with a change in Scripps score (3T: $\rho(18) = 0.58$, $p = 0.006$; ULF: $\rho(18) = 0.58$, $p = 0.006$).

ULF-detected PDA and number of EL were significantly associated with worsening Scripps scores over time ($\rho(18) = -0.48$, $p = 0.03$; $\rho(18) = -0.47$,

$p = 0.04$), but this association did not reach significance for 3T-derived PDA or DA ($\rho(18) = -0.29$, $p = 0.22$).

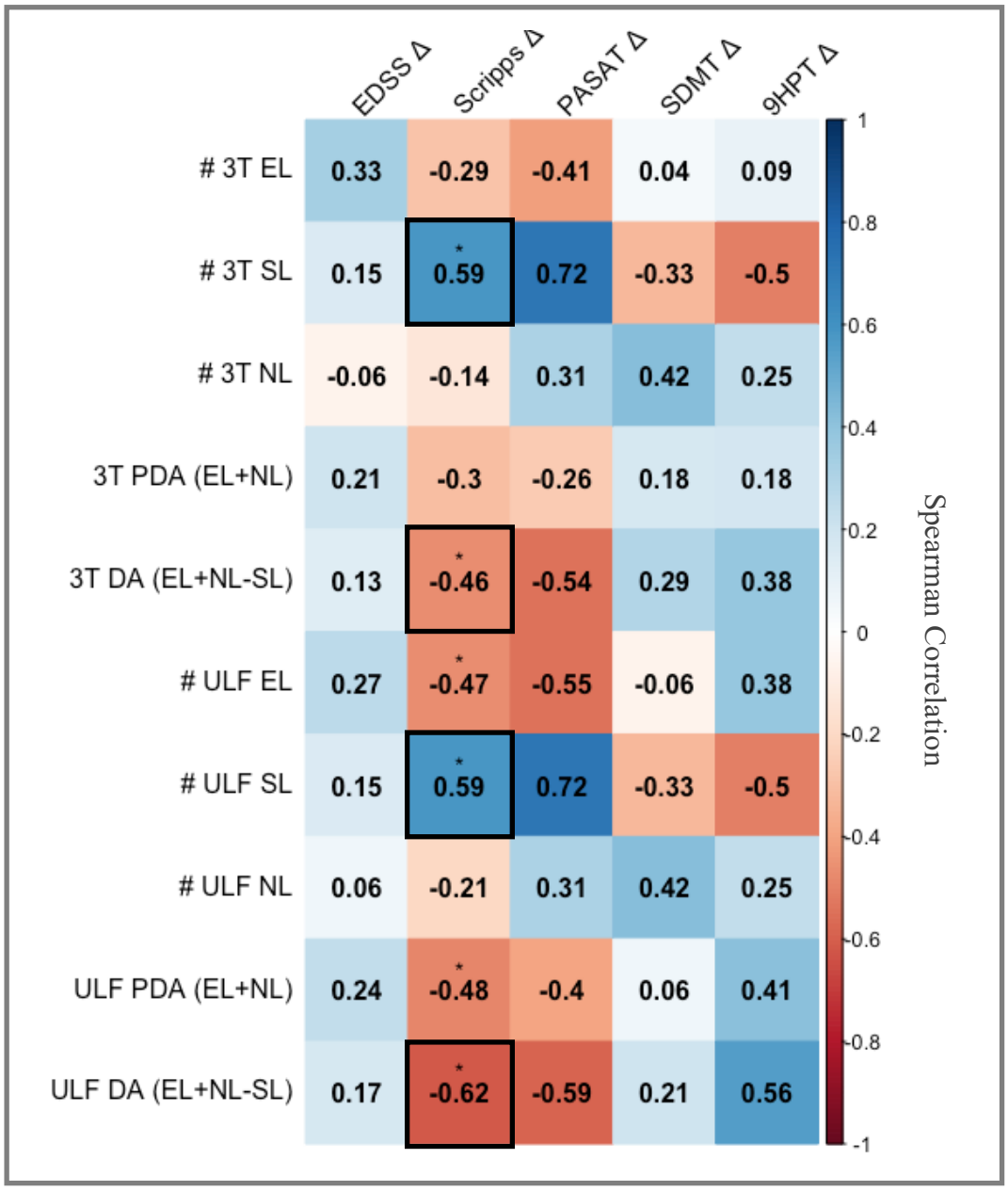


Figure 5.3: Spearman correlation matrix showing associations between change in clinical measures (Δ) and radiological lesion activity measures from 3T and ULF SUBLONG MRI. Significant correlations are marked with an asterisk (*). Black boxes highlight correlations that are significant for both ULF and 3T. Legend: EL=enlarging lesion; SL=shrinking lesion; NL=new lesion; PDA=positive disease activity; DA=disease activity; ULF=ultra-low field.

Figure 5.4A shows changes in clinical scores for participants with ULF case-level true positive (confirmed on 3T reference) PDA, NDA, and stable lesions. Among participants with PDA detected on ULF MRI, four showed

an increase in EDSS scores and three showed a decrease in Scripps scores, indicating worsening neurological function (clinical progression). Notably, the participants who exhibited these changes had longer follow-up intervals. Of the participants with PDA and available longitudinal 9HPT and PASAT data, one showed worsening on both tests, indicating declining motor and cognitive function. The other two showed mixed patterns of improvement and worsening. All three participants showed a slight increase in SDMT scores.

The one participant with case-level NDA detected by ULF demonstrated improvements in Scripps, PASAT, and 9HPT scores but a decline in SDMT and no change in EDSS. Participants with stable lesions (no NL, EL, or SL) showed mixed clinical trajectories, many of which had stable EDSS and Scripps.

Figure 5.4B illustrates clinical score changes in participants with ULF case-level false negatives. One participant with PDA, confirmed at 3T but missed at ULF, showed no change in EDSS or Scripps scores. In another case, ULF detected a shrinking lesion but missed an enlarging one. EDSS worsened slightly, while all other clinical scores showed mild improvement.

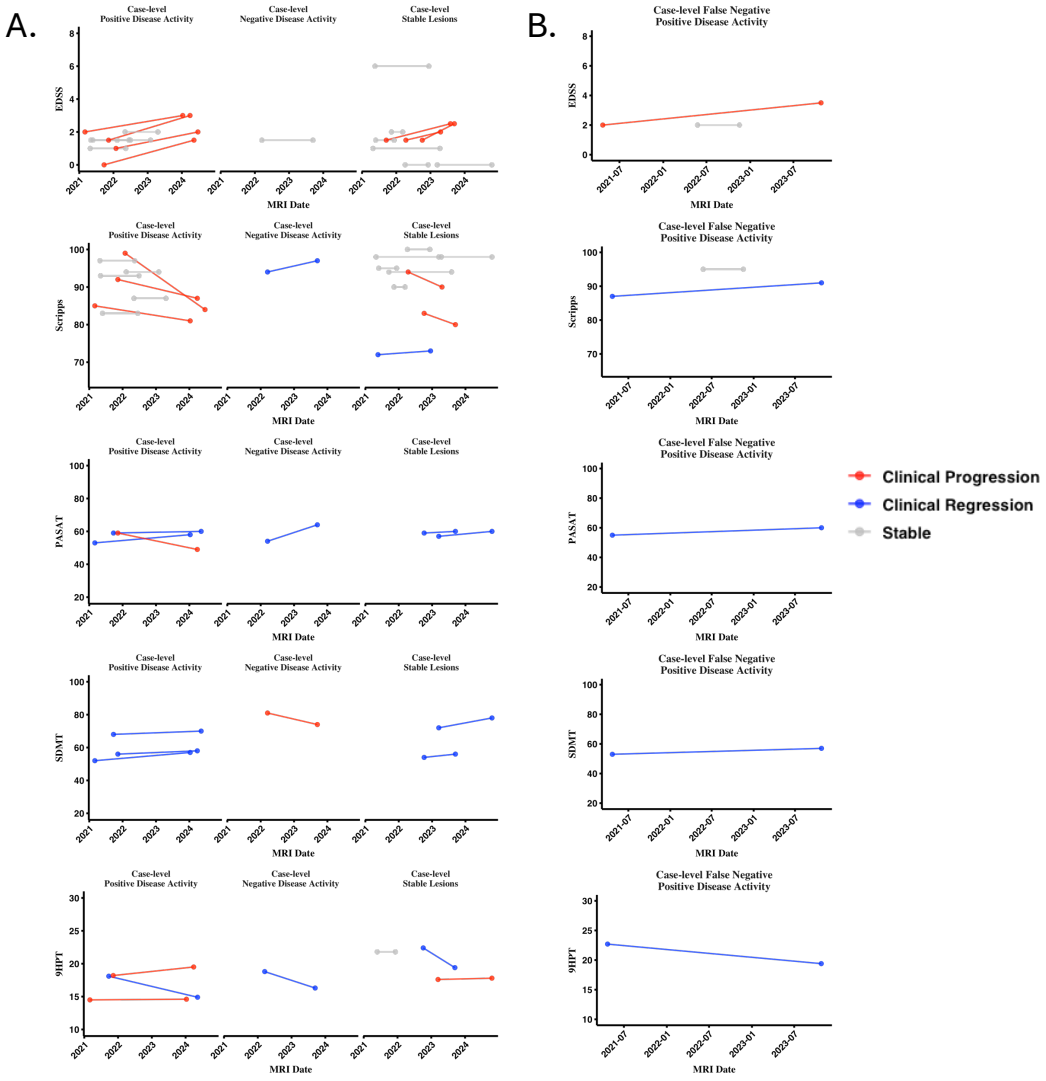


Figure 5.4: Longitudinal clinical score changes grouped by case-level ULF lesion activity detection. A) clinical score changes in participants with true-positive ULF assessments, grouped by case-level lesion activity: Positive Disease Activity (PDA), Negative Disease Activity (NDA), and Stable Lesions (definitions provided in text). B) Participants with false-negative PDA, where at least one new or enlarging lesion was present at 3T but none were detected at ULF. Lines are colour-coded: red indicates a change in the direction of clinical progression (worsening), blue indicates a change in the direction of clinical regression (improvement), and grey indicates no change in clinical score. These colours represent the slope of change in scores and may not directly reflect the patient's overall clinical status.

5.4 Discussion

In this chapter, preliminary evidence that ULF MRI-derived lesion measures may reflect clinically meaningful changes in MS was presented. Cross-sectionally, ULF-derived total WML volumes were associated with worsening EDSS and Scripps, using both Spearman correlations and age-adjusted linear models. In a smaller longitudinal subset, ULF-based assessments of new, enlarging, or shrinking lesions showed clinical associations capturing those observed at 3T.

These findings support the feasibility of ULF MRI as a tool for tracking WML activity relevant to patient outcomes.

Lesion burden categories are commonly used in clinical practice as descriptive markers of disease severity and serve as proxies for total WML volume by reflecting the overall extent of WML observed on MRI (Wattjes et al., 2021). In a cross-sectional dataset, total WML volumes differed significantly between lesion burden categories 1 (mild) and 2 (moderate), and between categories 1 and 3 (severe), using ULF and 3T segmentations. Prior findings in Chapter 4.1 demonstrated that total WML volume measures derived from ULF segmentations, both manual and automatic, correlate with 3T volumes despite showing a bias toward larger volumes at ULF. In this chapter, these results are extended by showing that ULF-derived lesion volumes successfully stratify patients according to lesion burden categories originally defined at 3T. No significant volume differences were observed between categories 2 and 3 at either field strength, likely reflecting limited sample sizes within these groups. However, lesion volumes generally increased with higher burden categories.

Clinical correlations between 3T and ULF total WML volume and disability were significant for both EDSS and Scripps scores in the cross-sectional analysis. In the longitudinal cohort, radiologically detected lesion activity correlated significantly with changes in Scripps scores, but not with changes in EDSS. The absence of significant correlations between radiologically detected lesion activity and changes in EDSS over time is unsurprising given the relatively short follow-up period, as average EDSS changes of around 0.5 points per year are common and may be difficult to detect within this time frame (Weinshenker, 1995). Additionally, both EDSS and Scripps were the most frequently available clinical scores across both cohorts, likely contributing to their stronger observed associations compared to other clinical measures with more limited data.

In the longitudinal cohort, several participants with case-level PDA detected at ULF showed little to no change in clinical disability measures over time. Conversely, some participants without detectable lesion evolution still experienced clinical worsening. This captures the well-described clinico-radiological paradox in MS, where there is not always a strong correlation between WML disease activity and the severity of a patient's clinical symptoms. This disconnect also reflects the complex nature of MS progression, where WML changes do not always correspond directly to clinical worsening, and factors like neu-

roplasticity, chronic inflammation, and neurodegeneration can independently impact patient outcomes (Kuhlmann et al., 2023; Portaccio et al., 2022).

The longitudinal correlations between radiological lesion activity and clinical disability presented here are preliminary and should be interpreted with caution due to the small sample size. No adjustments for covariates such as treatment status, baseline disability, or age were made, limiting the robustness of these findings. Despite these limitations, the primary purpose of this analysis was to provide initial evidence that ULF-derived metrics can capture clinical changes comparable to those detected with 3T imaging. These exploratory results lay the groundwork for future studies with larger cohorts and more rigorous statistical modelling that accounts for relevant confounders (similar to the statistical approaches for lesion modelling in Chapter 3.4.3).

In conclusion, this study leveraged paired 3T and ULF MRI data to show preliminary evidence that clinically relevant relationships established at 3T can be recapitulated at ULF. These findings support both the utility of the ULF-specific tools and pipelines introduced in this work and the broader potential of ULF MRI for tracking MS progression outside of conventional imaging environments. As ULF systems continue to improve in resolution, contrast, and analytic tooling, their capacity to support large-scale, accessible MS monitoring may expand significantly. Future research should build on these findings with larger, prospective cohorts to help define the role of ULF MRI in longitudinal MS care and research.

Chapter 6

Conclusions and Future Work

This thesis aims to address challenges in the detection and longitudinal characterization of white matter lesions (WMLs) in multiple sclerosis (MS) at non-standard MRI field strengths. WMLs are well visualized using conventional MRI (cMRI), which refers to clinically routine MRI sequences at 1.5T and 3T, making cMRI essential in MS clinical workflows and research. However, cMRI has limited specificity for chronic lesion pathology and is not always accessible in healthcare settings. Ultra-high field (UHF) and ultra-low field (ULF) MRI each offer advantages over conventional MRI (cMRI). UHF MRI is more sensitive to lesion microstructure and chronic active pathology, and ULF offers portable, low-cost imaging for tracking relevant WML disease activity. Despite these important advantages, methodological barriers have limited their use in WML studies. This work combines the development of dedicated WML analysis tools for UHF (7T) and ULF (64mT) with their application to unique datasets, including paired 3T-64mT scans in the same participants, a large longitudinal natural history cohort at 7T, and participants from two clinical trials scanned at 7T.

6.1 7T Tools and Applications

Tracking the evolution of chronic WMLs is critical to understanding progressive MS, where mechanisms like demyelination and chronic inflammation continue to drive disability even in the absence of new lesion formation (Cree et al., 2019; Hauser and Cree, 2020). Paramagnetic rim lesions (PRLs), which reflect iron-laden microglia at the lesion edge, are visible on susceptibility-sensitive imaging at 7T and represent a promising MRI biomarker of these chronic inflammatory processes, thought to underlie disease progression (Absinta et al.,

2016; Dal-Bianco et al., 2017). Quantitative T1 (qT1) mapping is a valuable tool for longitudinal assessment of lesion tissue destruction, as it is sensitive to demyelination and, at longer T1 times, also reflects axonal damage (Kolb et al., 2021). At 7T, this work focused on developing and evaluating image processing tools (Chapter 2.3.3) to enable assessment of chronic WML evolution and to support lesion-specific analyses of clinical trial interventions (Chapter 3.4.3).

To address the lack of image processing tools for studying chronic WMLs at 7T, a 3T-to-7T transfer learning segmentation model for whole-brain tissue and lesion segmentation, PLAn-7T, was developed and validated (Chapter 2.1). PLAn 7T outperformed existing segmentation models applied to 7T data in multiple assessments, including comparisons to manual lesion masks, neurologist consensus ratings, and 7T test-retest reproducibility. Unlike most segmentation pipelines developed for conventional MRI, PLAn 7T does not require FLAIR, which is often of suboptimal quality or unavailable at 7T (Saranathan et al., 2014), making it more broadly applicable across UHF studies. Both nnU-Net 7T and PLAn 7T deep learning models achieved strong performance using only MP2RAGE, demonstrating that this sequence provides sufficient lesion contrast to support accurate deep learning–based segmentation, consistent with recent findings (Fartaria et al., 2019). However, PLAn 7T provided specific improvements over nnU-Net 7T for WML segmentation, highlighting the benefits of 3T-to-7T transfer learning. PLAn-7T was pre-trained on 3T pseudo-labels and then fine-tuned on a limited 7T dataset. While transfer learning has previously improved MS lesion segmentation at 3T (Valverde et al., 2021), applying it across field strengths is novel and addresses the practical challenge of limited annotated data at 7T, especially where higher resolution makes manual labelling more tedious and time-consuming. These findings support the broader hypothesis that transfer learning can bridge domain shifts in high-resolution MRI, enabling scalable segmentation without the need for extensive hyper-parameter tuning or large labelled datasets.

The next stage of this work focused on developing a reliable longitudinal pipeline capable of tracking individual lesions from the PLAn 7T WML segmentation (Chapter 2.2). Pipeline choice strongly affected the reproducibility of lesion-level qT1 and volume measures, a factor that is rarely assessed in lesion-tracking studies. Pipelines with conservative masks yielded more stable qT1 values but were less sensitive to volume change, reflecting a trade-off between measurement stability and biological fidelity. Among the evaluated

approaches, a fully automated method with a lesion boundary correction step (Pipeline 5, Chapter 2.2) has a practical balance between fidelity and scalability. This pipeline achieved comparable reproducibility to a "gold-standard" manual-editing workflow. Prior 3T findings that median and 25th percentile qT1 values are robust statistics (Reich et al., 2015) were also confirmed in this work. These findings provide practical guidance for designing longitudinal lesion pipelines at 7T and establish a scalable framework for reliable lesion tracking.

The final tool developed for the 7T study of chronic WMLs is a lightweight, web-based Flask application for longitudinal visual assessment of lesion biomarkers, such as lesions with a paramagnetic rim (PRL or PRL+). Paramagnetic rims have been shown to have the capacity to resolve, making treatment of perilesional inflammation, as visualized on susceptibility-sensitive MRI, a promising target for clinical trials. PRL resolution is relatively rare, requiring researchers to review many lesions to identify the few that are resolving, making the process labour-intensive. The application enables efficient, blinded, and randomized lesion ratings across timepoints and participants, improving the ease of evaluation compared with traditional manual setups (Chapter 2.3). Additionally, the use of a highly compressed image format facilitates large-scale longitudinal studies while reducing storage requirements. The tool is open-source and reproducible, providing a flexible platform for multi-rater lesion evaluation. Beyond PRLs, it can be applied to other lesion-level biomarkers in MS, inter- and intra-rater reliability studies, and the creation of visually labelled training datasets.

The 7T tools developed in this work provide a foundation for lesion-specific analysis of chronic MS pathology. However, common limitations and insights emerged across these tools, highlighting areas for future research. First, these tools were designed for the use of clinical trial and natural history datasets acquired with specific protocols at the National Institutes of Health (NIH). This is particularly relevant for PLAn 7T and the boundary correction pipeline, which were developed for 7T MP2RAGE 0.7 mm isotropic scans. A key limitation is that these tools have not yet been evaluated across other sequences or datasets. Future work should validate them across multiple centres and acquisitions to enhance reproducibility and broader clinical applicability. While PLAn 7T is already available open-source, I plan to make the other tools openly available to support wider use. Despite these limitations, 7T MP2RAGE's widespread

use as a T1-weighted sequence at 7T makes it a logical and practical choice for developing 7T analysis tools.

Registration also emerged as an important consideration throughout the development of all 7T tools, influencing both design choices and performance. A key advantage of PLAn 7T was relying solely on a single contrast (MP2RAGE), thereby avoiding the complications of coregistering multiple sequences, which can be especially unreliable at 7T due to field-induced distortions. Longitudinal lesion tracking pipelines often rely on accurate registration across timepoints, but geometric distortions and subtle anatomical changes can reduce alignment quality. The boundary correction step in Pipeline 5 helped mitigate errors from misregistration by refining lesion borders to better align with the actual anatomy, thereby capturing both true biological changes and correcting registration-related segmentation errors. Additionally, the main limitation of the Flask application was its sensitivity to registration quality, which impacted rating reliability. Misalignments between timepoints or modalities could lead to incorrect lesion matching or inconsistent lesion appearance, potentially giving the false impression that lesions are changing when, in fact, different slices of the lesion are being displayed. Given these challenges, improving and standardizing registration methods at 7T remains an important area for future research to enhance the accuracy and reliability of lesion-specific analyses.

In addition to the technical evaluations detailed in Chapter 2.3.3, these tools were applied to 7T longitudinal datasets (Chapter 3.4.3), demonstrating their utility and scalability for 7T longitudinal studies while providing a dataset to evaluate their longitudinal performance further. For example, using a clinical trial dataset, the evaluation of the boundary correction pipeline (pipeline 5, Chapter 2.2) was extended beyond the test-retest timepoints and found that lesion segmentation variability and registration-related volume changes had minimal impact on longitudinal qT1 and volume trajectories. However, a key challenge that emerged in tracking WMLs in Chapter 3.4.3 was splitting the WML segmentation mask into individual lesions. In the boundary correction pipeline (Pipeline 5, Chapter 2.2), lesion masks were spatially separated into individual clusters. While effective in most cases, this approach struggled with confluent lesions or diffuse white matter (WM) damage, which were often merged into large spatial clusters that did not reflect biologically distinct lesions. This issue was much more common in the longitudinal dataset than in the small test-retest cohort, where only one participant exhibited lesions that

were difficult to separate. Accurate lesion separation is critical for tracking lesion-specific trajectories, as merged clusters can artificially inflate lesion size and distort qT1 measures.

Another insight from applying the 7T tools to longitudinal datasets was that, although the visualization application streamlined longitudinal lesion ratings, baseline PRL assessments remained labour-intensive. Standard guidelines for PRL detection require the paramagnetic rim to be visible around most of the lesion and on at least two slices or in multiple planes. In practice, the ability to scroll through the lesion across all three axes is necessary for PRL classification. To reduce the burden of baseline PRL ratings, future work should focus on building 3D-capable visualization tools within a rating application.

The longitudinal analysis of chronic WMLs in the 7T natural history study (Chapter 3.1) provides new insight into the evolution of PRL+ and PRL- lesions. PRL+ lesions had higher baseline qT1 values and volumes and showed steeper increases in qT1 over time, supporting their role as markers of ongoing damage. Tracking 81 participants for up to seven years confirmed, with robust longitudinal evidence, that PRL+ lesions are more destructive than PRL-. Their trajectory was consistent across periventricular, deep, and juxtacortical regions, and layer-specific qT1 analysis showed progressive damage at both core and rim. These findings underscore the clinical relevance of PRL+ lesions as therapeutic and monitoring targets.

In the ATaC-MS (anakinra) and BRaKE-MS (tolebrutinib) clinical trials, PRL+ lesion volumes and qT1 remained stable, suggesting a possible treatment effect given the worsening of these measures in the natural history data. However, in BRaKE-MS, the untreated reference group also showed no clear evidence of PRL+ worsening, suggesting that the absence of change may be due to limited sample size or an insufficient follow-up period to detect expected progression, although the reference group was not well-matched to the treatment group. Nevertheless, these findings highlight the heterogeneity in PRL+ lesion trajectories and the need to better understand why some PRL+ lesions in some patients exhibit more aggressive tissue damage than others, potentially by identifying lesion- and patient-level risk factors linked to their progression.

The findings from both the natural history and clinical trial data also revealed that the evolution of PRL- lesions is highly heterogeneous and warrants further study. In the natural history cohort, PRL- lesions generally increased in size, but qT1 decrease was mainly observed in deep white matter, rather than

in periventricular and juxtacortical regions suggesting potentially distinct underlying biological mechanisms. Similarly, PRL- lesions in the BRaKE-MS trial showed an unexpected trajectory, the biological significance of which remains unclear. These observations highlight gaps in our understanding of PRL- evolution heterogeneity. Future research should focus on identifying factors that influence PRL- capacity for remyelination, as this could have important implications for therapies targeting remyelination.

A limitation of this work is that the clinical trial and natural history datasets were analysed independently and in parallel, due to study-specific deadlines, so insights from one did not directly inform the other. For example, lesion location was not examined in detail in the clinical trial cohorts, which might have provided valuable insights into the behaviour of PRL- lesions. Integrating these datasets in future research could offer a more comprehensive understanding of lesion evolution. Since all these datasets remain available, there is potential for follow-up analyses to address these gaps.

The ATaC-MS and BRaKE-MS trials represent the first efforts to include change in paramagnetic rim in PRLs as imaging outcome measures in prospective interventional studies, marking an important step toward translating advanced 7T MRI biomarkers into the clinical trial setting. Although no PRL+ lesions showed complete rim resolution in either ATaC-MS or BRaKE-MS, the findings provide important considerations for future trials. Given that PRLs are known to persist for many years, complete resolution may not be a realistic short-term target. This underscores the need to determine whether to prioritize long-term tracking of PRL resolution or to assess if gradual changes, such as reduced rim intensity, are suitable short-term endpoints. Second, inter-rater agreement dropped when assessing longitudinal change in PRL, demonstrating the need for standardized rim-evolution guidelines and improved training resources. Additionally, quantitative evaluation of rim susceptibility using QSM may offer a more objective and sensitive approach to detecting gradual changes compared to qualitative visual assessment, potentially overcoming the inter-rater reliability challenges observed with longitudinal rim evaluation. However, given the small anatomical scale of rim regions and the impact of processing pipelines on lesion biomarkers demonstrated in this thesis (Chapter 2.2), robust validation of quantitative approaches would be essential before clinical implementation. Together, these insights highlight practical steps for strengthening the design and interpretation of PRL-focused therapeutic studies.

6.2 Ultra-Low Field MRI Tools and Applications

ULF MRI offers a promising, portable alternative to high-field imaging, but its lower resolution and lack of tailored analysis tools have limited its clinical use. This thesis addresses these challenges by developing and validating methods for reliable lesion assessment at ULF, comparing findings to paired 3T scans in people with MS.

To evaluate the feasibility of automated WML segmentation at ULF (Chapter 4.1), a comprehensive series of analyses were conducted that, for the first time, benchmark ULF segmentation performance using expert reference masks rather than algorithm-to-algorithm comparisons. Our results show that, despite known limitations in resolution and lesion detectability, particularly for small or posterior fossa lesions, ULF FLAIR imaging captures lesion volumes that correlate strongly with 3T, supporting its clinical feasibility. Building on this, I adapted the PLAn segmentation algorithm for ULF and found that it outperformed other methods, with FLAIR-only input being sufficient for robust WML segmentation. This is important given the practical need for short, simple ULF imaging protocols. ULF PLAn-derived lesion volumes also showed meaningful cross-sectional associations with disability measures, demonstrating clinical relevance (Chapter 5.3.1).

To explore the feasibility of using ULF MRI for tracking lesion dynamics over time, a longitudinal subtraction imaging pipeline was adapted to support radiological detection of evolving MS lesions. To our knowledge, this was the first longitudinal study of lesion change at ULF MRI. ULF reliably identified lesion enlargement and shrinkage with high case-level sensitivity, though detection of new lesions was more limited and depended on lesion size (Chapter 4.2). Compared to 3T, ULF tended to overestimate lesion volume and percent change, indicating a trade-off between sensitivity to change and volumetric precision. These findings contribute important new evidence that, although ULF may not yet be reliable for identifying small, new lesions needed for diagnosis, it holds promise for longitudinal follow-up and treatment monitoring, particularly in settings where higher field imaging is unavailable.

Finally, I explored whether combining orthogonal ULF images could enhance image quality, introducing a simple Fourier-based super-resolution method

(FouSR). ULF image quality is limited by resolution and noise. FouSR improved lesion boundary sharpness, offering a feasible, scanner-level alternative to deep learning methods that require disease-specific training. However, FouSR also introduced artifacts and is sensitive to motion and registration misalignment, highlighting the need for integrated correction methods like motion tracking and distortion mapping to realize its potential fully (Chapter 4.3).

Our ULF analyses highlight key challenges in lesion assessment in MS. Noisy normal-appearing white matter (WM) complicated the detection of small lesion changes and contributed to false positives, especially in the longitudinal subtraction pipeline and the FouSR method. Future work should prioritize advanced de-noising techniques to address this. Additionally, heterogeneity in scan protocols and software versions introduced variability, making it hard to separate scanner-related artefacts from inherent ULF limitations. While standardized acquisitions could reduce this variability, real-world ULF imaging is evolving rapidly, with frequent protocol changes aimed at improving image quality. To balance these issues, future efforts should focus on harmonization strategies tailored for ULF MRI.

The ULF studies presented in this thesis focused exclusively on participants with known or probable MS and did not include healthy controls or individuals with MS-mimicking conditions. As such, while longitudinal ULF MRI may capture lesion evolution consistent with dissemination in time, the absence of a control group limits conclusions about its diagnostic utility. Instead, the primary goal was to evaluate and improve the ability of ULF MRI to detect and characterize lesional changes over time. Given ULF's lower sensitivity for small or subtle lesions, its most promising clinical role may lie in follow-up imaging rather than initial MS diagnosis. When paired with a high-field MRI scan at baseline, ULF could provide sufficient sensitivity for routine monitoring of disease activity, potentially reducing costs and improving accessibility for patients in remote or resource-limited settings.

Importantly, although ULF MRI holds general promise for cost-effectiveness, the specific Hyperfine Swoop scanner used in this work incurs substantial costs, including subscription fees. Consequently, its primary advantages are portability, point-of-care use, and accessibility for patients with contraindications to higher-field MRI, rather than cost savings.

6.3 Concluding Thoughts on Ultra-High and Ultra-Low Field MRI in Multiple Sclerosis

UHF and ULF MRI play complementary roles in the visualization and monitoring of WML in MS. As demonstrated in this work, UHF MRI enables detailed characterization of lesion pathology through advanced biomarkers, such as PRLs, whose longitudinal evolution provides mechanistic insight and valuable clinical trial endpoints. These advanced markers help address the clinico-radiological paradox, where conventional lesion volume measurements often fail to fully explain clinical disability, and fill critical gaps in understanding progressive disease mechanisms in MS. This work also demonstrates that ULF-derived lesion volume and radiological changes, particularly at the patient level, are reasonably comparable to cMRI, supporting ULF's value for clinical follow-up, especially where high-field imaging is unavailable. This underscores the importance of advanced UHF imaging for the study of underlying disease processes while leveraging ULF's strengths for wider clinical use.

Despite their differences, common themes emerged across both field strengths when tracking WMLs:

- Insights from UHF and ULF tool development can inform each other. For example, flexible tools like PLAn, initially developed for UHF segmentation, were useful for ULF tasks as well, with pre-training on available 3T labels helping performance across both domains.

- Registration quality is critical, as accurate alignment of longitudinal images strongly impacts WML studies. At 7T, B0 inhomogeneity presents a significant challenge by causing distortions. At ULF, registration is essential not only for longitudinal tracking but also to enable comparison of ULF results with clinical standards. The Hyperfine ULF scanner exhibits coil sensitivity variations, especially in infratentorial regions, which can complicate both between-field-strength and longitudinal registration. These challenges underscore the need for carefully considered and tuned registration techniques and systematic evaluation of methods across non-standard field strengths.

- Validation in both controlled datasets and real-world clinical settings is essential. Tools must be tested not only in ideal conditions but also in relevant applications. For example, lesion splitting at 7T emerged as a

significant challenge in longitudinal datasets, despite being less of an issue during initial testing. At ULF, high noise levels in the WM complicated subtraction imaging, an issue not fully anticipated when adapting the subtraction pipeline, providing important feedback for future improvements.

Together, these points suggest a path forward where UHF and ULF MRI complement each other in MS lesion monitoring. They combine sensitive mechanistic biomarkers with accessible, scalable imaging. Tracking lesion-specific pathology across the full MRI spectrum can deepen our understanding of disease mechanisms to improve MS treatment and widen access to MRI monitoring for people living with MS.

Appendices

Appendix A

Dataset Details

This appendix describes the datasets and magnetic resonance imaging (MRI) acquisition parameters used throughout this thesis.

A.1 IRB Protocols

All data collection used in this thesis, including imaging, was performed at the National Institute of Neurological Disorders and Stroke (NINDS) in Bethesda, MD, USA, as part of the National Institutes of Health (NIH). The datasets fall under three Institutional Review Board (IRB) approved protocols (Table A.1). In all cases, participants provided written informed consent.

Table A.1: IRB Protocols for Datasets used in this thesis.

Protocol Name	**Title**	**Trial NCT**	**Protocol #**	**Description**	**Chpt. Ref.**	**App. Ref.**
Natural History	MRI to Evaluate Activity of MS	00001248	89-N-0045	Observational protocol with routine MRI at 3T, 64mT, and/or 7T in participants with MS or MS-mimics.	2 3 4 5	A.2 A.3 A.4 A.5
ATaC-MS	Anakinra for the Treatment of Chronically Inflamed WML in MS	04025554	19-N-0124	Interventional clinical trial targeting chronically inflamed WML in MS.	2 3	A.3
BRaKe-MS	Tolebrutinib, a Brain-Penetrant Bruton's Tyrosine Kinase Inhibitor, for Modulating Chronically Inflamed WML	04742400	21-N-0010	Interventional clinical trial targeting chronically inflamed WML in MS.	3	A.3

Abbreviations: MRI = magnetic resonance imaging; MS = multiple sclerosis; WML = white matter lesions; NCT = National Clinical Trial number

A.2 Datasets used in Chapter 2.3.3

This appendix section describes the datasets used in Chapter 2.3.3. Three subsets of data were derived from the Natural History Dataset Protocol (Table A.1). The MRI acquisition parameters for the sequences included in these datasets are provided in (Table A.2) and demographic of each subset are provided in (Table A.3).

1. Paired 7T and 3T Dataset. Dataset used in Chapter 2.1 for the evaluation of PLAn 7T. The *Paired 7T and 3T* Dataset consists of 40 participants who underwent both 7T and 3T scans within a maximum interval of 9 months, with a median interval of 36 days (range 1–96 days). The 40 participants were split into two cohorts Paired cohort 1 ($n = 25$) and Paired Cohort 2 ($n = 15$).

2. 7T Validation Cohort. This "7T Dataset" includes eight participants scanned exclusively at 7T used for the evaluation of PLAn 7T.

3. 7T Test-Retest Cohort. This dataset was used in Chapter 2.1 for the evaluation of PLAn 7T and in Chapter 2.2 for the evaluation of qT1 and volume reproducibility. The test-retest cohort consists of nine participants who underwent repeated 7T MP2RAGE scans. Each participant completed two 7T MP2RAGE scans during a single session, with a short break between scans during which the participant was removed from the scanner, the coil was detached, and the participant was instructed to reposition their head. One participant was excluded from analysis due to motion artefacts detected in the second scan.

Table A.2: MRI acquisition parameters for the datasets analysed in Chapter 2.3.3.

Scanner	Sequence	TR (ms)	TE (ms)	TI (ms)	Flip Angle (°)	Voxel Size	Coil
3T Skyra	3D T1w MP2RAGE	5000	2.26	700, 2500	4.5	1mm iso	32-ch
	3D FLAIR	5000	393	1800	—	1mm iso	32-ch
	2D PD/T2 FSE	3630	9.6, 96	—	90/180*	0.7×0.7×3	32-ch
7T Magnetom	3D T1w MP2RAGE	4000	4.6	350, 1350	4.5	0.7mm iso	1-ch Tx, 32-ch Rx

Abbreviations: TR = repetition time; TE = echo time; TI = inversion time; iso = isotropic; FSE = fast spin echo; Tx = transmit; Rx = receive. *Refocusing flip angles were variable (nominal 180°).

Table A.3: Participant Demographics for Datasets analysed in Chapter 2.3.3. Continuous data shown as mean ± SE

	Paired 7T-3T Cohort		**7T Validation Cohort**	**7T Test-Retest Cohort**
	Paired Cohort 1	Paired Cohort 2		
# Participants (n)	25	15	8	9
Age (y)	54.9 ± 3.3	54.5 ± 2.2	68.6 ± 2.2	61.9 ± 4.5
Sex (M,F)	7, 18	3, 12	4, 4	3, 6
Sub-Diagnosis *(RRMS, SPMS, PPMS, RIS, OND)*	12, 10, 1, 1, 1	10, 5, 0, 0, 0	3, 4, 1, 0, 0	4, 2, 1, 1, 1
Race *(White, Black, Other*)*	22, 1, 2	11, 4, 0	7, 2, 0	6, 2, 0, 1
Ethnicity *(NH/Lat, H/Lat, Other*)*	22, 2, 1	15, 0, 0	7, 1, 0	9, 0, 0

Abbreviations: y = years; M = male; F = female; RRMS = relapsing-remitting multiple sclerosis; SPMS = secondary progressive multiple sclerosis; PPMS = primary progressive multiple sclerosis; OND = other neurological disease ; NH/Lat = not Hispanic/Latinx ; H/Lat= Hispanic/Latinx. *Information not provided or marked as other

A.3 Datasets used in Chapter 3.4.3

This appendix provides MRI acquisition details for the datasets analysed in Chapter 3.4.3. The 7T acquisition parameters used across all sections of the chapter are summarized in Table A.4. Relevant acquisitions included sagittal 3D MP2RAGE (0.7-mm isotropic) and gradient recalled echo T2*-weighted (T2*-GRE) magnitude and phase images acquired in three partially overlapping volumes to achieve near-complete supratentorial brain coverage. Examples of concatenated *HIRES* and *AMRI* T2*-GRE magnitude and phase images are shown in Figure A.1. Further details regarding *AMRI* T2*-GRE acquisitions with and without navigator-based motion correction can be found in the original publication (Liu, Beck, Filippini, van Gelderen, de Zwart, Norato, Sati, Al-Louzi, Kolb, Donadieu, Morrison, Duyn and Reich, 2021).

Table A.4: MRI acquisition parameters for the datasets analysed in Chapter 2.3.3.

Sequence	TR (ms)	TE (ms)	TI (ms)	Flip Angle (°)	Voxel Size (mm)
3D T1w MP2RAGE	4000	4.6	350, 1350	4, 5	$0.7 \times 0.7 \times 0.7$
2D T2*-GRE †	1300	15, 32	-	50	$0.2 \times 0.2 \times 1$
2D T2*-GRE ‡	4095	11.4, 22.5, 33.6, 44.7, 55.8	-	90	$0.5 \times 0.5 \times 0.5$
3D MoCo B0Co T2*-GRE ‡	74	18, 29.5, 41.0, 52.4	-	10	$0.5 \times 0.5 \times 0.5$

Abbreviations: TR = repetition time; TE = echo time; TI = inversion time; iso = isotropic; T2*-GRE = T2*-weighted gradient recalled echo; MoCo B0Co T2*-GRE = navigator-guided motion and B0 corrected T2*-weighted gradient recalled echo
† referred to as *HIRES* T2*-GRE in text
‡ referred to as *AMRI* T2*-GRE in text

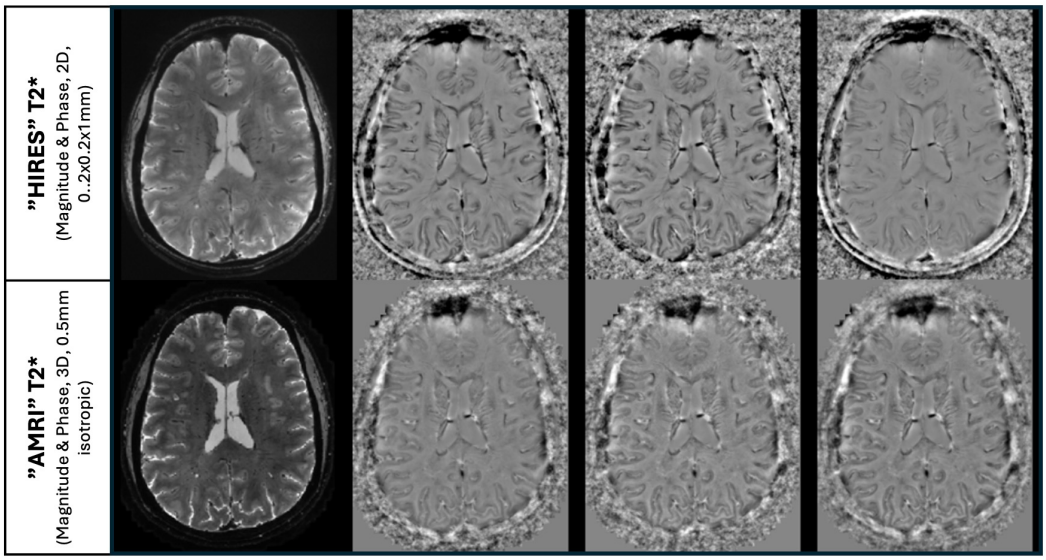


Figure A.1: Examples of *HIRES* (top) and concatenated *AMRI* (bottom) acquisitions. From left to right: averaged magnitude, unwrapped averaged phase, and individual unwrapped phase echoes.

A.3.1 Dataset used in Chapter 3.1

Longitudinal MP2RAGE data were acquired on a 7T whole-body research system (Siemens) equipped with a single-channel transmit and 32-channel phased-array receive head coil (Table A.4, first row). Software versions included *syngo_MR_B17* ($n = 51$) and *syngo_MR_E12* ($n = 76$).

At baseline, participants underwent T2*-GRE imaging. Fifty-seven participants had *HIRES* acquisitions (Table A.4, second row), and 23 participants

had *AMRI* (Table A.4, third/fourth row), three of whom were scanned with navigator motion correction (Table A.4, fourth row).

A.3.2 Dataset used in Chapter 3.2

Longitudinal MP2RAGE data were acquired on the same 7T system with the single-channel transmit and 32-channel phased-array head coil (Table A.4, first row), using software version *syngo_MR_E12*. Longitudinal T2*-GRE data were acquired with the *HIRES* protocol (Table A.4, second row).

A.3.3 Dataset used in Chapter 3.3

Longitudinal MP2RAGE data were acquired on the same 7T system and coil setup (Table A.4, first row), using software version *syngo_MR_E12*.

Longitudinal T2*-GRE data were acquired using both *HIRES* and navigator-corrected *AMRI*. For the analyses presented in the main text, the navigator-corrected *AMRI* acquisitions were used (Table A.4, last row).

A.4 Datasets used in Chapter 4.3.4

This appendix section describes the paired ULF-3T datasets used in Chapter 4.3.4. All data were acquired under the Natural History Protocol (Table A.1). HF MRI was acquired on a 3T Siemens Skyra scanner using either a 20- or 32-channel head coil. ULF MRI scans were acquired on the portable 64mT Hyperfine Swoop MRI scanner.

A.4.1 Dataset used in Chapter 4.1

This retrospective dataset included 104 participants with known or suspected MS who underwent paired ULF and HF MRI scans on the same day between April 2021 and June 2024. An overview of the MRI acquisition parameters are provided in Table A.5.

Table A.5: Summary of MRI acquisition parameters for paired HF (3T) and ULF (64mT) scans used in Chapter 4.1. ULF values represent the minimum and maximum for each parameter across all participants ($n = 104$).

MRI Sequence	**TR (s)**	**TE (s)**	**TI (s)**	**Flip Angle (°)**	**Slice thick-ness (mm)**	**In-plane voxel size (mm^2)**
64 mT 3D T2-FLAIR (Ax.)	3.5–4	0.14–0.21	1.3–1.4	90	5-6	1.7–2.0
64 mT 3D T1w (Ax.)	0.9–1.5	0.005–0.006	0.3–0.4	90	5	1.6
3T 3D T1w-MP2RAGE	5	0.0028	0.7/2.5	4.5	1	1
3T 3D T2-FLAIR	4.8	0.35	1.8	120	1	1

Abbreviations: TR = repetition time; TE = echo time; TI = inversion time; Ax = Axial

Acquisition parameters, reconstruction methods (Cartesian or deep learning), and software versions varied across ULF T1w and T2-FLAIR sequences. Table A.6 summarises acquisitions grouped by common parameters, reporting the total participants per group and their distribution between training and validation subsets.

Table A.6: Grouped ULF MRI acquisition parameters for Chapter 4.1, showing total participants per group and training/validation split.

MRI Sequence	TR (s)	TE (s)	TI (s)	Flip Angle (°)	Recon.	Soft. Ver.	Slice thickness (mm)	In-plane voxel size (mm^2)	Part. (n)	Training/ Validation Split
64 mT 3D T2-FLAIR (Ax.)	3.5	0.16	1.3	90	cart	rc8.7.0_Beta1	5	1.6	1	0/1
	3.5	0.16	1.3	90	cart	rc8.7.0Beta1	5	1.7	48	9/39
	4	0.18	1.4	90	cart	8.3.0	5	1.8	4	1/3
	4	0.18	1.4	90	cart	8.3.2	5	1.8	4	1/3
	4	0.18	1.4	90	cart	rc8.3.2	5	1.8	3	1/2
	4	0.19	1.4	90	cart	rc8.3.2	5	1.6	1	0/1
	4	0.2	1.2	90	cart	rc8.5.0	5	1.8	4	2/2
	4	0.21	1.4	90	cart	rc8.5.0	5	1.6	4	1/3
	4	0.14	1.4	90	cart	rc8.6.0	6	2	1	0/1
	4	0.17	1.4	90	cart	rc8.6.0	5	1.7	12	2/10
	4	0.21	1.4	90	cart	rc8.7.0Beta1	5	1.6	1	0/1
	4	0.19	1.4	90	DL	8.2.0	5	1.6	4	0/4
	4	0.19	1.4	90	DL	rc8.2.0	5	1.6	13	3/10
	4	0.18	1.4	90	DL	rc8.3.0	5	1.8	3	0/3
	4	0.2	1.4	90	DL	rc8.3.0	5	1.6	1	0/1
64 mT 3D T1w (Ax.)	0.9	0.006	0.4	90	cart	rc8.5.0	5	1.6	1	0/1
	0.9	0.005	0.3	90	cart	rc8.7.0Beta1	5	1.6	47	8/39
	1	0.006	0.3	90	cart	rc8.6.0	5	1.6	1	0/1
	1.5	0.005	0.3	90	cart	8.3.0	5	1.6	4	1/3
	1.5	0.005	0.3	90	cart	8.3.2	5	1.6	4	1/3
	1.5	0.005	0.3	90	cart	rc8.3.2	5	1.6	4	1/3
	1.5	0.005	0.3	90	cart	rc8.5.0	5	1.6	7	3/4
	1.5	0.005	0.3	90	cart	rc8.6.0	5	1.6	12	2/10
	1.5	0.005	0.3	90	cart	rc8.7.0Beta1	5	1.6	3	1/2
	1.5	0.006	0.3	90	DL	8.2.0	5	1.6	4	0/4
	1.5	0.006	0.3	90	DL	rc8.2.0	5	1.6	13	3/10
	1.5	0.005	0.3	90	DL	rc8.3.0	5	1.6	4	0/4

Abbreviations: TR = repetition time; TE = echo time; TI = inversion time; Ax = Axial; Sag = Sagittal; Recon = Reconstruction Method; Soft. Ver. = Software Version; DL = Deep Learning; cart = Cartesian; Part. = Participants; n = number;

A.4.2 Dataset used in Chapter 4.2

Baseline and follow-up MRI data were collected for 23 participants with known or suspected MS. At each timepoint, paired ULF and HF MRI scans were acquired on the same day, including FLAIR sequences. An overview of the FLAIR acquisition parameters is provided in Table A.7, and participant characteristics

are summarised in Table A.8.

ULF acquisition parameters were heterogeneous, particularly with respect to software version and image reconstruction method (deep learning–based or Cartesian). ULF acquisition varied between participants and within participants across timepoints, and is discussed in relation to the results of Chapter 4.2 in Appendix D.2.2.

Table A.7: Summary of MRI acquisition parameters for paired HF (3T) and ULF (64mT) scans used in Chapter 4.2. ULF values represent the minimum and maximum for each parameter across all participants ($n = 23$) at both baseline and follow-up.

MRI Sequence	**TR (s)**	**TE (s)**	**TI (s)**	**Flip Angle (°)**	**Slice thickness (mm)**	**In-plane voxel size (mm^2)**
64 mT 3D T2-FLAIR (Ax.)	3.5–4	0.16–0.21	1.3–1.4	90	5	1.6–1.8
3T 3D T2-FLAIR	4.8	0.35	1.8	120	1	1

Abbreviations: TR = repetition time; TE = echo time; TI = inversion time; Ax = Axial

Table A.8: Participant Characteristics for Dataset used in Chapter 4.2. Data shown as count or mean ± SE, with ranges [min, max] where applicable.

Participant Characteristics	**All Participants**
Baseline Age in years	50.5 ± 2.1
Sex (Female, Male)	17, 6
Sub-Diagnoses (RRMS, PPMS, SPMS, RIS)	19, 2, 1, 1
Race (White, Black, Asian)	19, 3, 1
Follow-Up Period (months)	16.9 ± 1.9

Abbreviations: SE = standard error; RRMS = relapsing remitting MS; SPMS = secondary progressive MS; PPMS = primary progressive MS; RIS = Radiologically Isolated Syndrome

A.4.3 Dataset used in Chapter 4.3

MRI data were collected for 10 participants with known or suspected MS and in a standard Hyperfine phantom. Paired ULF and HF MRI scans were acquired on the same day, including FLAIR and T1w acquisitions. At ULF, two ULF FLAIR images were acquired in the axial and coronal direction (software version: rc8.6.0 and Cartesian reconstruction method). An overview of the FLAIR

acquisition parameters is provided in Table A.9, and participant characteristics are summarised in Table A.10.

Table A.9: Summary of MRI acquisition parameters for paired HF (3T) and ULF (64mT) scans used in Chapter 4.3.

MRI Sequence	**TR (s)**	**TE (s)**	**TI (s)**	**Flip Angle (°)**	**Slice thick-ness (mm)**	**In-plane voxel size (mm^2)**
64 mT 3D T2-FLAIR (Ax.)	4	0.17	1.4	90	5	1.7
64 mT 3D T2-FLAIR (Cor.)	4	0.17	1.4	90	5	1.7
3T 3D T2-FLAIR	4.8	0.35	1.8	120	1	1
3T 3D T1w-MP2RAGE	5	0.0028	0.7/2.5	4.5	1	1

Abbreviations: TR = repetition time; TE = echo time; TI = inversion time; Ax = Axial; Sag = Sagittal

Table A.10: Participant Characteristics for Dataset used in Chapter 4.3. Data shown as count or mean ± SE, with ranges [min, max] where applicable.

		All Participants (n=10)
Participant Demographics	Age (years)	44.1 ± 4.1
	Sex (Female, Male)	6, 4
	Sub-Diagnoses (RRMS, probable MS)	8, 2
	Race (White, Black, Other)	6, 3, 1
Whole Brain MRI Characteristics	Total 3T WML Volume (mm^3)	4313 ± 10026

Abbreviations: MS = multiple sclerosis; RRMS = relapsing-remitting MS; WML = white matter lesion; SE = standard error

A.5 Datasets used in Chapter 5.4

This appendix describes the paired ULF-3T datasets used in Chapter 5.4, a subset with MS diagnosis of those in Chapter 4.3.4. All data were collected under the Natural History Protocol (Table A.1). HF MRI was acquired on a 3T Siemens Skyra using a 20- or 32-channel head coil, and ULF MRI on a 64mT Hyperfine Swoop scanner.

The MRI acquisition details for the dataset used in cross-sectional analyses (methods described in Section 5.2.1 of the main text) are provided in Table A.11. Demographic and participant characteristics are provided in the main text.

Table A.11: Summary of MRI acquisition parameters for cross sectional paired HF (3T) and ULF (64mT) scans used in Chapter 5.4. ULF values represent the minimum and maximum for each parameter across all participants ($n = 72$).

MRI Sequence	**TR (s)**	**TE (s)**	**TI (s)**	**Flip Angle (°)**	**Slice thickness (mm)**	**In-plane voxel size (mm^2)**
64 mT 3D T2-FLAIR (Ax.)	3.5–4	0.14–0.21	1.3–1.4	90	5-6	1.7–2.0
64 mT 3D T1w (Ax.)	0.9–1.5	0.005–0.006	0.3–0.4	90	5	1.6
3T 3D T1w-MP2RAGE	5	0.0028	0.7/2.5	4.5	1	1
3T 3D T2-FLAIR	4.8	0.35	1.8	120	1	1

Abbreviations: TR = repetition time; TE = echo time; TI = inversion time; Ax = Axial

The MRI acquisition details for the dataset used in longitudinal analyses (methods described in Section 5.2.2 of the main text) are provided in Table A.12. Demographic and participant characteristics are provided in the main text.

Table A.12: Summary of MRI acquisition parameters for longitudinal paired HF (3T) and ULF (64mT) scans used in Chapter 5.4. ULF values represent the minimum and maximum for each parameter across all participants ($n = 21$) at both baseline and follow-up.

MRI Sequence	**TR (s)**	**TE (s)**	**TI (s)**	**Flip Angle (°)**	**Slice thickness (mm)**	**In-plane voxel size (mm^2)**
64 mT 3D T2-FLAIR (Ax.)	3.5–4	0.16–0.21	1.3–1.4	90	5	1.6–1.8
3T 3D T2-FLAIR	4.8	0.35	1.8	120	1	1

Abbreviations: TR = repetition time; TE = echo time; TI = inversion time; Ax = Axial

Appendix B

Supplementary Data for Chapter 2.3.3

This appendix provides supplementary data and materials supporting the experimental work presented in Chapter 2.3.3.

B.1 Supplementary Data for Chapter 2.1

This appendix section contains additional analyses from Chapter 2.1, which describes the development and evaluation of an automatic 7T whole-brain and lesion segmentation algorithm using a single MRI modality. Figure B.1 compares all overlap metrics calculated using FSL's *bianca_overlap_metrics* between nnU-Net 7T and PLAn-7T, with the manual 7T WML segmentation serving as the reference. Across all overlap metrics in the 7T validation dataset, PLAn-7T consistently outperformed nnU-Net 7T.

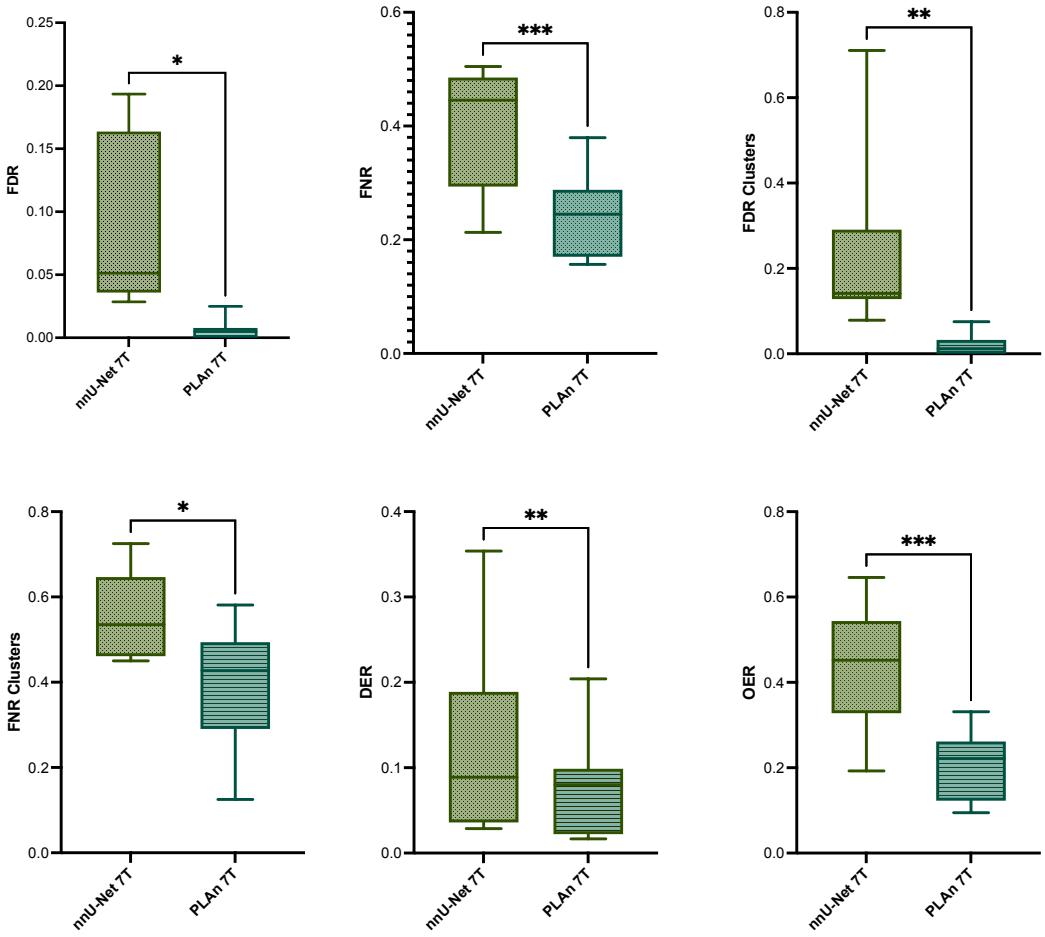


Figure B.1: PLAn-7T outperforms nnU-Net 7T across all Bianca overlap metrics for lesion segmentation. Overlap metrics were calculated from the 7T validation dataset using the manual 7T WML segmentation as the reference.

B.2 Supplementary Data for Chapter 2.2

This appendix section includes supplementary information from Chapter 2.2, which examines the impact of longitudinal image analysis pipelines on lesion quantitative T1 values and volume measurements. Tables B.1 and B.1 display all bias and limits of agreement for Bland-Altman analyses conducted and plotted in the main text.

Table B.1: Bias and limits of agreement (LoA) across pipelines for tissue-level measures referenced in Chapter 2.2, Figure 2.9

Measure	**Tissue**	**Pipeline**	**Bias**	**Lower LoA**	**Upper LoA**	**Figure**
Median qT1	GM	Pipeline 1	10.3	-14.1	26.4	Figure 2.9A
Median qT1	GM	Pipeline 2	9.1	-16.9	18.9	Figure 2.9A
Median qT1	GM	Pipeline 3	11.4	-22.6	20.5	Figure 2.9A
Median qT1	GM	Pipeline 4	13.4	-22.2	30.5	Figure 2.9A
Median qT1	WM	Pipeline 1	3.3	-8.2	6.5	Figure 2.9A
Median qT1	WM	Pipeline 2	2.9	-6.3	5.3	Figure 2.9A
Median qT1	WM	Pipeline 3	3.0	-7.0	5.6	Figure 2.9A
Median qT1	WM	Pipeline 4	3.3	-7.4	5.6	Figure 2.9A
Median qT1	CSF	Pipeline 1	168.3	50.8	286	Figure 2.9A
Median qT1	CSF	Pipeline 2	16.4	49.8	82.5	Figure 2.9A
Median qT1	CSF	Pipeline 3	20.1	-35.7	75.9	Figure 2.9A
25th percentile qT1	GM	Pipeline 5	4.5	-48.1	57.1	Figure 2.9B
Median qT1	GM	Pipeline 5	4.2	-22.0	30.5	Figure 2.9B
75th percentile qT1	GM	Pipeline 5	1.5	-40.9	49.6	Figure 2.9B
Mean qT1	GM	Pipeline 5	0.1	-35.7	59.5	Figure 2.9B
25th percentile qT1	WM	Pipeline 5	-4.4	-6.5	3.8	Figure 2.9B
Median qT1	WM	Pipeline 5	-1.3	-7.6	5.0	Figure 2.9B
75th percentile qT1	WM	Pipeline 5	-1.3	-8.9	6.4	Figure 2.9B
Mean qT1	WM	Pipeline 5	-1.9	-9.1	5.4	Figure 2.9B
25th percentile qT1	CSF	Pipeline 5	15.4	-43.3	74.1	Figure 2.9B
Median qT1	CSF	Pipeline 5	5.5	-42.3	53.3	Figure 2.9B
75th percentile qT1	CSF	Pipeline 5	12.9	-29.7	55.5	Figure 2.9B
Mean qT1	CSF	Pipeline 5	20.1	-21.2	61.5	Figure 2.9B

Abbreviations: LoA = limits of agreement; qT1 = quantitative T1 time; GM = grey matter; WM = white matter; CSF = cerebral spinal fluid

Table B.2: Bland-Altman Statistics for Lesion Analysis Stages 1 and 2 Referenced in Chapter 2.2, Figures 2.10 and 2.12

Measure	Pipeline	Bias	Lower LoA	Upper LoA	Analysis	Figure
Median qT1	Pipeline 1	7.8	-113.2	128.8	Lesion Analysis Stage 1	Figure 2.10A
Median qT1	Pipeline 2	13.3	-100.6	127.2	Lesion Analysis Stage 1	Figure 2.10A
Median qT1	Pipeline 3	1.9	-97.3	101.2	Lesion Analysis Stage 1	Figure 2.10A
Median qT1	Pipeline 4	17.0	-94.0	127.9	Lesion Analysis Stage 1	Figure 2.10A
Median qT1	Pipeline 5	4.7	-105.2	114.5	Lesion Analysis Stage 1	Figure 2.10A
Volume	Pipeline 1	N/A	0.5	8.2	Lesion Analysis Stage 1	Figure 2.10B
Volume	Pipeline 2	N/A	0.8	7.4	Lesion Analysis Stage 1	Figure 2.10B
Volume	Pipeline 3	N/A	2.0	6.3	Lesion Analysis Stage 1	Figure 2.10B
Volume	Pipeline 4	N/A	-27.5	-33.4	Lesion Analysis Stage 1	Figure 2.10B
Volume	Pipeline 5	N/A	0.6	7.2	Lesion Analysis Stage 1	Figure 2.10B
25th percentile qT1	Pipeline 5	-2.5	-102.4	97.5	Lesion Analysis Stage 1	Figure 2.10C
Median qT1	Pipeline 5	-1.6	-100.6	97.5	Lesion Analysis Stage 1	Figure 2.10C
75th percentile qT1	Pipeline 5	-11.6	-109.4	86.2	Lesion Analysis Stage 1	Figure 2.10C
Mean qT1	Pipeline 5	-9.1	-105.2	87.1	Lesion Analysis Stage 1	Figure 2.10C
Median qT1	Pipeline 1	5.81	-68.8	80.5	Lesion Analysis Stage 2	Figure 2.12A
Median qT1	Pipeline 2	5.41	-69.8	80.6	Lesion Analysis Stage 2	Figure 2.12A
Median qT1	Pipeline 3	9.3	-64.2	82.9	Lesion Analysis Stage 2	Figure 2.12A
Median qT1	Pipeline 4	6.9	-61.3	75.1	Lesion Analysis Stage 2	Figure 2.12A
Median qT1	Pipeline 5	6.9	-61.6	75.5	Lesion Analysis Stage 2	Figure 2.12A
Volume	Pipeline 1	-35.4	-252	182.3	Lesion Analysis Stage 2	Figure 2.12B
Volume	Pipeline 2	N/A	N/A	N/A	Lesion Analysis Stage 2	Figure 2.12B
Volume	Pipeline 3	N/A	-280.1	282.7	Lesion Analysis Stage 2	Figure 2.12B
Volume	Pipeline 4	N/A	-270.8	270.9	Lesion Analysis Stage 2	Figure 2.12B
Volume	Pipeline 5	N/A	-280.1	282.7	Lesion Analysis Stage 2	Figure 2.12B
25th percentile qT1	Pipeline 5	-0.1	-60.6	60.4	Lesion Analysis Stage 2	Figure 2.12C
Median qT1	Pipeline 5	7.4	-103.9	118.7	Lesion Analysis Stage 2	Figure 2.12C
75th percentile qT1	Pipeline 5	9.8	-96.9	54.5	Lesion Analysis Stage 2	Figure 2.12C
Mean qT1	Pipeline 5	12.7	-92	117.3	Lesion Analysis Stage 2	Figure 2.12C

Abbreviations: LoA = limits of agreement; qT1 = quantitative T1 time

Appendix C

Supplementary data for Chapter 3.4.3 .

C.1 Supplementary data for Chapter 3.1

C.1.1 Summary of Covariates Considered for Linear Mixed Models

This section outlines the covariates evaluated and included in the linear mixed effects models (LMMs) presented in Chapter 3.1, which aimed to quantify longitudinal changes in chronic lesion volume and qT1 using a 7T natural history dataset.

C.1.1.1 Baseline Linear Mixed Effects Models

Models 1, 2, 5, and 6 from Chapter 3.1, Table 3.1 assessed baseline differences in lesion volume or qT1 as a function of PRL status. The following covariates were considered:

1. Age at MRI scan (continuous): Included to control for age-related changes in MS progression
2. Sex (categorical: Female/Male): Included due to known sex-related differences in MS disease expression and progression.
3. Treatment efficacy (categorical: High, Moderate, Low, Untreated): To account for the potential impact of disease-modifying (DMT) therapies on WML characteristics.
4. Presence of enhancing lesions at baseline (Yes/No): Acute inflammation may alter qT1 values and lesion morphology.

5. Scanner covariate (6 categories): Encodes different 7T scanners and hardware/software combinations (i.e., *FMRIFD7TsyngoMRB17*) to adjust for non-biological variance imaging biomarkers.

6. Normalized median qT1 of the skull-stripped whole brain: Used to account for session-related scanner fluctuations. Although qT1 is a quantitative measure, variability due to coil loading or scanner drift can influence global values. Whole-brain qT1 was preferred over normal appearing white matter qT1 due to better model fit (data not shown) and reduced sensitivity to MS-related changes in normal-appearing tissue.

7. Approximate age at onset: Initially considered but excluded due to collinearity with age at scan.

C.1.1.2 Longitudinal Linear Mixed Effects Models

Models 3, 4, 7, and 8 from Chapter 3.1, Table 3.1 examined changes in lesion volume or qT1 over time. Covariates were largely consistent with baseline models, with several additions:

1. Sex, treatment efficacy, scanner covariate, and normalized median whole-brain qT1: Carried over from baseline models.

2. Presence of enhancing lesion at any timepoint (Yes/No): Included to capture dynamic inflammatory activity over follow-up.

3. Presence of new lesions at any timepoint (Yes/No): Controls for varying levels of overall disease activity (e.g., one participant had ≥ 15 new lesions).

4. Approximate age at onset (continuous): Retained here to help disentangle age-related and disease-duration-related effects.

5. Total follow-up time (continuous): Included to account for variability in scan intervals that may affect the magnitude of observed change.

C.1.1.3 Longitudinal Linear Mixed Effects Models: Lesion Layers

Models 9 and 10 from Chapter 3.1, Table 3.1 evaluated qT1 in inner vs outer lesion layers and included the same covariates as the corresponding baseline and longitudinal models. One additional covariate was added:

1. Normalized lesion volume at baseline: Larger lesions may affect the spatial meaning of lesion layers (e.g., in small lesions, layer 2 may fall within the core, whereas in large lesions, it may reflect peripheral tissue)

C.1.2 Sensitivity Analyses for LMMs in Chapter 3.1

This appendix summarizes sensitivity analyses for the LMMs described in Chapter 3.1.

For models with substantial skew or kurtosis in the residuals, outliers were defined as observations with standardized studentized residuals $\geq |3|$. These observations were removed and the models were refit (Sensitivity Analysis 1). This analysis was performed only for models showing skewness or kurtosis.

A second sensitivity analysis excluded voxels affected by partial voluming in WML segmentations (red voxels in Figure 3.1; Sensitivity Analysis 2). Since partial voluming voxels were already excluded in the lesion layer analyses (Section 3.1.2.2), this analysis was not applied to LMM 9 and 10 (Table 3.1).

All sensitivity analyses retained the same covariates and model design as the main models. Main and sensitivity models are reported for comparison. Only key effects of interest are shown.

Quantitative WML Analysis: PRL+ vs PRL-. Residuals from baseline qT1 and volume LMMs (Models 1 and 2, Table 3.1) did not show substantial skew or kurtosis. Therefore, Sensitivity Analysis 1 was not performed. Table C.1 reports the fixed effect coefficients for PRL and the estimated marginal means for PRL+ and PRL– lesions in the main models and Sensitivity Analysis 2 (partial voluming voxels removed). The findings for both models remained consistent between the main models and sensitivity analysis models.

Table C.1: Sensitivity analysis 2 for baseline qT1 and volume LMM: PRL+ vs PRL-. Fixed effect coefficients and predicted group values (estimated marginal means) for PRL+ and PRL– lesions in main models and sensitivity analysis 2 (partial voluming voxels removed).

Model	Norm. Dependent Var.	Effect/Group*	Main Model (Estimate, SE, p)	Sensitivity 2 (Estimate, SE, p)
1	qT1	PRL	$0.03 \pm 0.003, < 0.001$	$0.4 \pm 0.02, < 0.001$
		PRL+	$6.5 \pm 0.03, < 0.001$	$6.4 \pm 0.03, < 0.001$
		PRL-	$6.1 \pm 0.02, < 0.001$	$6.0 \pm 0.02, < 0.001$
2	Volume	PRL	$1.3 \pm 0.06, < 0.001$	$1.4 \pm 0.07, < 0.001$
		PRL+	$4.8 \pm 0.07, < 0.001$	$4.8 \pm 0.07, < 0.001$
		PRL-	$3.5 \pm 0.04, < 0.001$	$4.5 \pm 0.08, < 0.001$

Abbreviations: Norm. Dependent Var = normalized dependent variable; qT1 = quantitative T1 time; PRL = paramagnetic rim lesion; SE = standard error; p = p-value
**Effect* refers to LMM fixed effect (PRL), *Group* refers to estimated marginal means (least-squares means) for each PRL group.

Residuals of longitudinal qT1 and volume LMMs (Models 3 and 4 from Table 3.1) exhibited kurtosis. Tables C.2 and C.3 present the fixed interaction effects (PRL × age) and the estimated slopes for each PRL group (PRL+ and PRL–) for the main model as well as Sensitivity Analyses 1 and 2. Results were consistent across the main models and both sensitivity analyses.

Table C.2: Sensitivity analysis 1 for longitudinal qT1 and volume LMM: PRL+ vs PRL-. Fixed effect coefficients and predicted group slopes for PRL+ and PRL– lesions in main models and sensitivity analysis 1 (exclusion of statistical outliers).

Model	Norm. Dependent Var.	Effect/Group*	Main Model (Estimate, SE, p)	Sensitivity 1 (Estimate, SE, p)
3	qT1	PRL×age	$0.009 \pm 0.002, < 0.001$	$0.008 \pm 0.02, < 0.001$
		PRL+ slope	$0.007 \pm 0.004, 0.04$	$0.006 \pm 0.003, 0.04$
		PRL- slope	$-0.002 \pm 0.004, 0.6$	$-0.003 \pm 0.003, 0.3$
4	Volume	PRL×age	$0.006 \pm 0.005, 0.02$	$0.005 \pm 0.04, 0.03$
		PRL+ slope	$0.02 \pm 0.005, < 0.001$	$0.02 \pm 0.004, < 0.001$
		PRL- slope	$0.02 \pm 0.002, < 0.001$	$0.02 \pm 0.003, < 0.001$

Abbreviations: Norm. Dependent Var = normalized dependent variable; qT1 = quantitative T1 time; PRL = paramagnetic rim lesion; SE = standard error; p = p-value
**Effect* refers to LMM fixed interaction term (PRL×age), *Group* refers to estimated slope for each PRL group.

Table C.3: Sensitivity analysis 2 for longitudinal qT1 and volume LMM: PRL+ vs PRL-. Fixed effect coefficients and predicted group slopes for PRL+ and PRL– lesions in main models and sensitivity analysis 2 (partial voluming voxels removed).

Model	**Norm. Dependent Variable**	**Effect/Group**	**Main Model (Estimate, SE, p)**	**Sensitivity 2 (Estimate, SE, p)**
3	qT1	PRL×age	$0.009 \pm 0.002, < 0.001$	$0.01 \pm 0.02, < 0.001$
		PRL+ slope	$0.007 \pm 0.004, 0.04$	$0.009 \pm 0.004, 0.02$
		PRL- slope	$-0.002 \pm 0.004, 0.6$	$-0.003 \pm 0.003, 0.3$
4	Volume	PRL×age	$0.006 \pm 0.005, 0.02$	$-0.002 \pm 0.005, 0.7$
		PRL+ slope	$0.02 \pm 0.005, < 0.001$	$0.02 \pm 0.004, < 0.001$
		PRL- slope	$0.02 \pm 0.002, < 0.001$	$0.02 \pm 0.003, < 0.001$

Abbreviations: Norm. Dependent Var = normalized dependent variable; qT1 = quantitative T1 time; PRL = paramagnetic rim lesion; SE = standard error; p = p-value
**Effect* refers to LMM fixed interaction term (PRL×age), *Group* refers to estimated slope for each PRL group.

Quantitative WML Analysis: Lesion Location. Residuals from the baseline location qT1 and volume LMMs (Models 5 and 6, Table 3.1) did not indicate substantial skew or kurtosis; therefore, Sensitivity Analysis 1 was not performed. LMM contrasts by PRL and lesion location for the main models and Sensitivity Analysis 2 (partial voluming voxels removed) are provided in Table C.4. PRL+ vs PRL– differences in both qT1 and volume and within-PRL group qT1 differences between lesion locations remained consistent across main and sensitivity analyses. In contrast, within-PRL group volume contrasts that were significant in the main models (all DWM vs. JC, DWM vs. PV, and JC vs. PV contrasts in PRL– lesions, and the DWM vs. PV contrast in PRL+ lesions) were no longer significant after exclusion of partial voluming voxels. This suggests that the location-dependent differences observed in the main analysis may have been driven by partial volume errors.

Table C.4: Sensitivity analysis 2 for baseline qT1 and volume LMM contrasts by PRL and lesion location. Main model and Sensitivity 2 results show estimated contrasts (Δ), standard errors (SE), and p-values.

Model	Norm. Dependent Var.	Group	Contrast	Main Model (Δ, SE, p)	Sensitivity 2 (Δ, SE, p)
		DWM		-0.3 ± 0.05, $< .0001$	-0.3 ± 0.05, $< .0001$
5	qT1	JC	PRL- vs. PRL+	-0.2 ± 0.05, $< .0001$	-0.3 ± 0.05, $< .0001$
		PV		-0.3 ± 0.04, $< .0001$	-0.5 ± 0.04, $< .0001$
			DWM vs. JC	-0.4 ± 0.02, $< .0001$	-0.2 ± 0.02, $< .0001$
5	qT1	PRL-	DWM vs. PV	-0.4 ± 0.02, $< .0001$	-0.2 ± 0.02, $< .0001$
			JC vs. PV	-0.04 ± 0.02, 0.07	-0.03 ± 0.02, 0.3
			DWM vs. JC	-0.3 ± 0.07, 0.0001	-0.19 ± 0.06, 0.005
5	qT1	PRL+	DWM vs. PV	-0.5 ± 0.06, $< .0001$	-0.4 ± 0.06, $< .0001$
			JC vs. PV	-0.2 ± 0.06, 0.004	-0.2 ± 0.06, 0.0002
		DWM		-1.14 ± 0.12, $< .0001$	-1.16 ± 0.13, $< .0001$
6	Volume	JC	PRL- vs. PRL+	-0.92 ± 0.12, $< .0001$	-1.03 ± 0.14, $< .0001$
		PV		-1.24 ± 0.09, $< .0001$	-1.26 ± 0.11, $< .0001$
			DWM vs. JC	-0.48 ± 0.05, $< .0001$	0.03 ± 0.06, 0.85
6	Volume	PRL-	DWM vs. PV	-0.31 ± 0.04, $< .0001$	-0.01 ± 0.05, 0.98
			JC vs. PV	0.17 ± 0.05, 0.001	-0.04 ± 0.06, 0.77
			DWM vs. JC	-0.26 ± 0.16, 0.24	0.17 ± 0.18, 0.62
6	Volume	PRL+	DWM vs. PV	-0.42 ± 0.15, 0.012	-0.11 ± 0.17, 0.78
			JC vs. PV	-0.16 ± 0.14, 0.50	-0.28 ± 0.16, 0.21

Abbreviations: Norm. Dependent Var = normalized dependent variable; qT1 = quantitative T1 time; PRL = paramagnetic rim lesion; SE = standard error; p = p-value; estimated contrasts (Δ)

Residuals of longitudinal qT1 and volume LMMs by PRL and location (Models 7 and 8 from Table 3.1) exhibited kurtosis. Tables C.5 and C.6 show estimate slopes for PRL/Location groups and contrasts by PRL and lesion location for the main model as well as Sensitivity Analyses 1 and 2. Results were consistent across the main models and both sensitivity analyses.

Table C.5: Sensitivity analysis 1 for longitudinal qT1 and volume LMM contrasts by PRL and lesion location. Main model and Sensitivity 1 results show estimated slopes or estimated contrasts (Δ), standard errors (SE), and p-values.

Model	Norm. Dependent Var.	Contrast	Group	Main Model (Δ, SE, p)	Sensitivity 2 (Δ, SE, p)
7	qT1	PRL-	DWM	-0.012 ± 0.003, < 0.001	-0.01 ± 0.003, 0.002
		PRL+	DWM	0.001 ± 0.006, 0.908	-0.001 ± 0.005, 0.82
		PRL-	JC	0.004 ± 0.003, 0.233	-0.004 ± 0.004, 0.31
		PRL+	JC	0.01 ± 0.005, 0.061	0.009 ± 0.005, 0.07
		PRL-	PV	-0.005 ± 0.003, 0.098	-0.004 ± 0.003, 0.31
		PRL+	PV	0.002 ± 0.005, 0.656	0.008 ± 0.005, 0.09
7	qT1	PRL- vs PRL+	DWM	-0.01 ± 0.01, 0.01	-0.011 ± 0.004, 0.016
			JC	-0.01 ± 0.005, 0.18	-0.006 ± 0.004, 0.17
			PV	-0.01 ± 0.004, 0.05	-0.006 ± 0.003, 0.05
7	qT1	DWM vs. JC	PRL-	-0.02 ± 0.002, < 0.001	-0.015 ± 0.002, < 0.001
		DWM vs. PV		-0.007 ± 0.002, < 0.001	-0.006 ± 0.001, < 0.001
		JC vs. PV		0.009 ± 0.002, < 0.001	0.009 ± 0.002, < 0.001
		DWM vs. JC	PRL+	-0.009 ± 0.007, 0.33	-0.01 ± 0.006, 0.2
		DWM vs. PV		-0.001 ± 0.006, 0.97	-0.001 ± 0.005, 0.98
		JC vs. PV		0.008 ± 0.006, 0.34	0.009 ± 0.005, 0.2
8	volume	PRL-	DWM	0.03 ± 0.003, < 0.001	0.02 ± 0.003, < 0.001
		PRL+	DWM	0.03 ± 0.01, 0.002	0.03 ± 0.009, 0.002
		PRL-	JC	0.01 ± 0.003, < 0.001	0.009 ± 0.003, 0.002
		PRL+	JC	0.03 ± 0.009, 0.002	0.02 ± 0.008, 0.002
		PRL-	PV	0.02 ± 0.003, < 0.001	0.02 ± 0.003, < 0.001
		PRL+	PV	0.03 ± 0.007, < 0.001	0.03 ± 0.006, < 0.001
8	volume	PRL- vs PRL+	DWM	-0.002 ± 0.01, 0.83	-0.002 ± 0.01, 0.9
			JC	-0.02 ± 0.01, 0.09	-0.01 ± 0.01, 0.07
			PV	-0.01 ± 0.01, 0.45	-0.004 ± 0.01, 0.5
8	volume	DWM vs. JC	PRL-	0.014 ± 0.003, < 0.001	0.015 ± 0.003, < 0.001
		DWM vs. PV		0.007 ± 0.003, 0.06	0.004 ± 0.003, 0.3
		JC vs. PV		-0.007 ± 0.003, 0.05	-0.01 ± 0.003, < 0.001
		DWM vs. JC	PRL+	0.001 ± 0.01, 0.99	0.002 ± 0.01, 0.98
		DWM vs. PV		0.004 ± 0.011, 0.94	0.001 ± 0.01, 0.99
		JC vs. PV		0.003 ± 0.011, 0.97	-0.001 ± 0.01, 0.99

Abbreviations: Norm. Dependent Var = normalized dependent variable; qT1 = quantitative T1 time; PRL = paramagnetic rim lesion; SE = standard error; p = p-value; estimated contrasts (Δ)

Table C.6: Sensitivity analysis 2 for longitudinal qT1 and volume LMM contrasts by PRL and lesion location. Main model and Sensitivity 2 results show estimated slopes or estimated contrasts (Δ), standard errors (SE), and p-values.

Model	Norm. Dependent Var.	Contrast	Group	Main Model (Δ, SE, p)	Sensitivity 2 (Δ, SE, p)
7	qT1	PRL-	DWM	-0.012 ± 0.003, < 0.001	-0.01 ± 0.003, 0.002
		PRL+	DWM	0.001 ± 0.006, 0.908	-0.001 ± 0.005, 0.8
		PRL-	JC	0.004 ± 0.003, 0.233	-0.004 ± 0.004, 0.3
		PRL+	JC	0.01 ± 0.005, 0.061	0.009 ± 0.005, 0.07
		PRL-	PV	-0.005 ± 0.003, 0.098	-0.004 ± 0.003, 0.3
		PRL+	PV	0.002 ± 0.005, 0.656	0.008 ± 0.005, 0.09
7	qT1	PRL- vs PRL+	DWM	-0.01 ± 0.01, 0.01	-0.009 ± 0.004, 0.02
			JC	-0.01 ± 0.005, 0.18	-0.01 ± 0.004, 0.001
			PV	-0.01 ± 0.004, 0.05	-0.01 ± 0.003, 0.0009
7	qT1	DWM vs. JC	PRL-	-0.02 ± 0.002, < 0.001	-0.01 ± 0.002, < 0.001
		DWM vs. PV		-0.007 ± 0.002, < 0.001	-0.01 ± 0.001, < 0.001
		JC vs. PV		0.009 ± 0.002, < 0.001	-0.0001 ± 0.002, 0.99
		DWM vs. JC	PRL+	-0.009 ± 0.007, 0.33	-0.01 ± 0.005, 0.13
		DWM vs. PV		-0.001 ± 0.006, 0.97	-0.01 ± 0.005, 0.18
		JC vs. PV		0.008 ± 0.006, 0.34	0.001 ± 0.005, 0.95
8	volume	PRL-	DWM	0.03 ± 0.003, < 0.001	0.03 ± 0.003, < 0.001
		PRL+	DWM	0.03 ± 0.01, 0.002	0.03 ± 0.01, 0.009
		PRL-	JC	0.01 ± 0.003, < 0.001	0.02 ± 0.004, < 0.001
		PRL+	JC	0.03 ± 0.009, 0.002	0.008 ± 0.01, 0.4
		PRL-	PV	0.02 ± 0.003, < 0.001	0.01 ± 0.004, 0.001
		PRL+	PV	0.03 ± 0.007, < 0.001	0.03 ± 0.009, 0.001
8	volume	PRL- vs PRL+	DWM	-0.002 ± 0.01, 0.83	0.0001 ± 0.01, 0.99
			JC	-0.02 ± 0.01, 0.09	0.01 ± 0.01, 0.2
			PV	-0.01 ± 0.01, 0.45	-0.02 ± 0.01, 0.07
8	volume	DWM vs. JC	PRL-	0.014 ± 0.003, 0	0.006 ± 0.004, 0.3
		DWM vs. PV		0.007 ± 0.003, 0.06	0.015 ± 0.004, < 0.001
		JC vs. PV		-0.007 ± 0.003, 0.05	0.009 ± 0.004, 0.11
		DWM vs. JC	PRL+	0.001 ± 0.01, 1	0.018 ± 0.01, 0.4
		DWM vs. PV		0.004 ± 0.011, 0.94	-0.001 ± 0.01, 0.99
		JC vs. PV		0.003 ± 0.011, 0.97	-0.019 ± 0.01, 0.3

Abbreviations: Norm. Dependent Var = normalized dependent variable; qT1 = quantitative T1 time; PRL = paramagnetic rim lesion; SE = standard error; p = p-value; estimated contrasts (Δ)

Quantitative WML Analysis: Lesion Layer. Residuals from the baseline and longitudinal layer qT1 and volume LMMs (Models 9 and 10, Table 3.1) showed no evidence of substantial skewness or kurtosis; therefore, Sensitivity Analysis 1 was not performed. Partial volume voxels had already been excluded in the main analysis, as detailed in the methods section (Figure 3.1).

C.2 Supplementary Data for Chapter 3.2

Table C.7: Inclusion and Exclusion Criteria for ATaC-MS Clinical Trial

Inclusion Criteria	**Exclusion Criteria**
• Age $\geq$ 18 • Ability to give informed consent • If fertile, agreement to use an effective method of birth control during the study and for up to 3 months after the last dose of the study drug • Agreement not to participate in any other interventional study while participating in this protocol • Diagnosis of MS, either stable or clinically progressive, with evidence of no new lesion formation by comparison of baseline MRI scan with a historical MRI scan at least 2 years prior • Prior 7-tesla MRI scan, with high image quality in the judgement of the study neuroradiologist, demonstrating at least one white matter lesion with a paramagnetic rim	• Pregnancy or current breastfeeding • Use of another investigational agent within 1 month of screening • Active infection and/or neutropenia (ANC $<$ 1000 cells/μL) • History of lymphoma • Known hypersensitivity to administration of anakinra • Previous treatment with anakinra and/or TNF-receptor inhibitor • History of asthma • QuantiFERON-TB gold positive • Prior treatment with anti-CD20 agent (ocrelizumab, rituximab) • Prior treatment with anti-CD52 agent (alemtuzumab) • History or current evidence of any condition, therapy, or laboratory abnormality that might confound the results of the trial or interfere with participation for the full duration of the trial • Renal dysfunction, as defined by Clinical Center guidelines for administration of gadolinium • Liver dysfunction, as indicated by baseline AST or ALT $>$ 1.5 times the upper limit of normal • Clinical relapse or new lesion formation in the 12 months prior to dosing • One or more gadolinium-enhancing lesions on the screening scan • Change in disease-modifying therapy in the 6 months prior to dosing • Medical contraindication for 7-tesla MRI (e.g., non-organic implant, cardiac pacemaker, infusion pump, metallic implants, objects, or body piercings that are not MRI-compatible or cannot be removed) • Psychological contraindication for 7-tesla MRI (e.g., claustrophobia) • Contraindication to gadolinium administration • Active neoplastic disease or any medical condition, other than MS, that requires concurrent immunosuppression or immunomodulation

Six lesions that were originally classified as PRL+ in the baseline PRL ratings, were reclassified as *unratable* (not true PRL+) during longitudinal analysis: 5 lesions in longitudinal PRL rating 1, 1 lesion in longitudinal PRL rating 2 (Figure C.1).

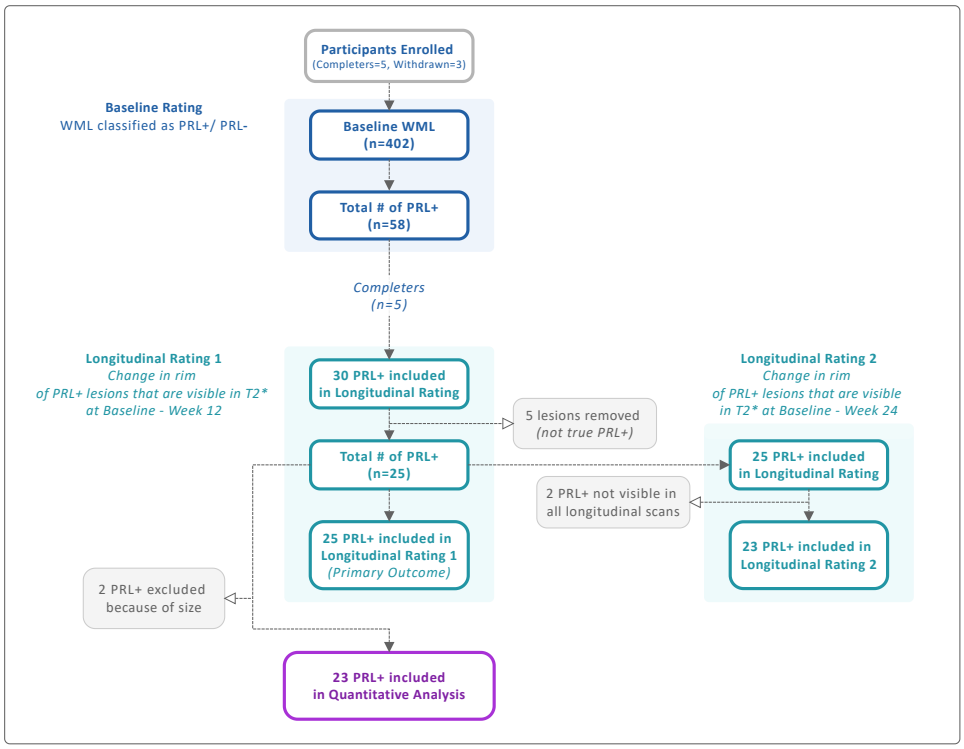


Figure C.1: An overview of PRL+ lesions included in various analyses in the ATaC-MS clinical trial.

Overview of PRL+ lesions included in ATaC-MS trial analyses, including reclassification of six baseline PRL+ lesions during longitudinal review.

Table C.8: Post-hoc Pairwise Comparisons from RM-ANCOVA of Normalized Lesion qT1 (Full Dataset with Imputed Missing Data)

PRL	Time 1	Time 2	Estimate	SE	DF	Tukey-Kramer Adj. p-value
PRL-	0	4	0.02932	0.009173	811	0.0637
	0	8	0.02263	0.009519	870	0.4228
	0	12	0.02952	0.009468	820	0.0797
	0	16	0.01764	0.008852	766	0.6985
	0	24	0.02156	0.009038	887	0.4173
	4	8	-0.00669	0.009518	841	0.9999
	4	12	0.000203	0.009405	872	1.0000
	4	16	-0.01168	0.008945	892	0.9784
	4	24	-0.00776	0.009324	769	0.9996
	8	12	0.006893	0.009685	846	0.9999
	8	16	-0.00499	0.009362	833	1.0000
	8	24	-0.00107	0.009635	853	1.0000
	12	16	-0.01188	0.009241	881	0.9808
	12	24	-0.00797	0.009610	796	0.9996
	16	24	0.003916	0.009006	725	1.0000

Residual diagnostics for the qT1 RM-ANCOVA model revealed non-normality of residuals, with Shapiro-Wilk statistics for studentized and Pearson residuals indicating significant deviation from normality ($W = 0.9887$ and 0.99,

both $p < 0.0001$), as well as moderate skewness (-0.34) and elevated kurtosis (0.72). To examine the impact of extreme values, a sensitivity analysis was conducted by excluding 9 outlier observations (three single-timepoint lesions and one lesion measured at all six timepoints). This exclusion modestly improved normality ($W = 0.99$ and 0.99; $p < 0.0001$), reducing skewness (-0.18) and kurtosis (0.26).

Table C.9: Post-hoc Pairwise Comparisons from RM-ANCOVA of Normalized Lesion qT1 after removing statistical outliers, 8 timepoints (Full Dataset with Imputed Missing Data)

PRL	Time 1	Time 2	Estimate	SE	DF	t-value	Tukey-Kramer Adj. p-value
PRL-	0	4	0.03320	0.008901	876	3.73	0.0110
	0	8	0.02476	0.009226	862	2.68	0.2363
	0	12	0.03319	0.009191	848	3.61	0.0168
	0	16	0.01892	0.008534	785	2.22	0.5380
	0	24	0.02311	0.008824	825	2.62	0.2707
	4	8	-0.00845	0.009336	856	-0.90	0.9991
	4	12	-0.00001	0.009295	867	0.00	1.0000
	4	16	-0.01429	0.008795	842	-1.62	0.9000
	4	24	-0.01010	0.009130	810	-1.11	0.9944
	8	12	0.00843	0.009496	850	0.89	0.9992
	8	16	-0.00584	0.009137	802	-0.64	1.0000
	8	24	-0.00165	0.009422	848	-0.18	1.0000
	12	16	-0.01427	0.009080	825	-1.57	0.9189
	12	24	-0.01008	0.009408	829	-1.07	0.9957
	16	24	0.00419	0.008835	688	0.47	1.0000

Table C.10: Post-hoc Pairwise Comparisons from RM-ANCOVA of Normalized Lesion Volume (Full Dataset with Imputed Missing Data)

PRL	Timepoint 1	Timepoint 2	Estimate	SE	DF	t-value	Tukey-Kramer Adj. p-value
PRL-	0	4	0.01332	0.02980	1021	0.45	1
	0	8	0.01825	0.03080	1022	0.59	1
	0	12	-0.00106	0.03080	1022	-0.03	1
	0	16	0.10200	0.02870	1021	3.55	0.0203
	0	24	0.07661	0.02870	1021	2.67	0.2433
	4	8	0.00493	0.03132	1021	0.16	1
	4	12	-0.01438	0.03132	1021	-0.46	1
	4	16	0.08869	0.02980	1021	2.98	0.1171
	4	24	0.06329	0.02980	1021	2.12	0.6057
	8	12	-0.01931	0.03189	1021	-0.61	1
	8	16	0.08376	0.03080	1022	2.72	0.2178
	8	24	0.05836	0.03080	1022	1.90	0.7623
	12	16	0.10310	0.03080	1022	3.35	0.0401
	12	24	0.07767	0.03080	1022	2.52	0.3271
	16	24	-0.02540	0.02870	1021	-0.89	0.9993

Table C.11: Post-hoc Pairwise Comparisons from RM-ANCOVA of Normalized Lesion qT1 (Excluding participants with Imputed Data)

PRL	Time 1	Time 2	Estimate	SE	DF	t-value	Tukey-Kramer Adj. p-value
	0	4	0.03829	0.009319	728	4.11	0.0026
	0	8	0.02755	0.009336	739	2.95	0.1255
	0	12	0.03590	0.009324	708	3.85	0.0071
	0	16	0.02257	0.009313	690	2.42	0.3915
	0	24	0.02823	0.009349	761	3.02	0.1049
	4	8	-0.01074	0.009201	772	-1.17	0.9912
	4	12	-0.00239	0.009170	782	-0.26	1.0000
PRL-	4	16	-0.01572	0.009148	772	-1.72	0.8593
	4	24	-0.01007	0.009230	757	-1.09	0.9950
	8	12	0.00835	0.009197	784	0.91	0.9990
	8	16	-0.00498	0.009178	778	-0.54	1.0000
	8	24	0.00068	0.009249	776	0.07	1.0000
	12	16	-0.01333	0.009143	790	-1.46	0.9511
	12	24	-0.00767	0.009225	775	-0.83	0.9996
	16	24	0.00566	0.009210	756	0.61	1.0000

Table C.12: Post-hoc Pairwise Comparisons from RM-ANCOVA of Normalized Lesion volume (Excluding participants with Imputed Data)

PRL	Time 1	Time 2	Estimate	SE	DF	t-value	Tukey-Kramer Adj. p-value
	0	4	0.03637	0.03011	731	1.21	0.9883
	0	8	0.03546	0.03010	732	1.18	0.9905
	0	12	0.01318	0.03009	729	0.44	1.0000
	0	16	0.12670	0.03080	752	4.11	0.0025
	0	24	0.08045	0.03003	672	2.68	0.2386
	4	8	-0.00091	0.02959	770	-0.03	1.0000
	4	12	-0.02319	0.02958	772	-0.78	0.9998
PRL-	4	16	0.09029	0.03064	675	2.95	0.1270
	4	24	0.04408	0.02939	762	1.50	0.9406
	8	12	-0.02228	0.02956	768	-0.75	0.9998
	8	16	0.09120	0.03064	671	2.98	0.1176
	8	24	0.04499	0.02937	760	1.53	0.9315
	12	16	0.11350	0.03064	668	3.70	0.0122
	12	24	0.06727	0.02936	762	2.29	0.4836
	16	24	-0.04621	0.03068	610	-1.51	0.9388

C.3 Supplementary Data for Chapter 3.3

Table C.13: Inclusion and Exclusion Criteria for BRaKE-MS Clinical Trial Treatment Cohort

Inclusion Criteria	**Exclusion Criteria**
• Able to provide informed consent • Willingness to comply with all study procedures and availability for the duration of the study • Male or female, aged ≥ 18 • Diagnosed with multiple sclerosis according to the 2017 revision of the McDonald diagnostic criteria, with no new lesion formation by comparison of baseline MRI scan with a historical MRI scan at least 6 months prior • On intravenous anti-CD20 antibody treatment (specifically ocrelizumab or rituximab) for at least 6 months, with the most recent dose at most 3 months prior to enrolment. • Willing to forego further anti-CD20 antibody treatment for the duration of the study • Has a prior 7-tesla MRI scan, no more than 1 year prior to enrolment, demonstrating at least one white matter lesion with a paramagnetic rim • For females of reproductive potential: agrees to use highly effective contraception for at least 1 month prior to dosing and to use such a method during study participation and for an additional 12 weeks after the end of tolebrutinib administration • For males of reproductive potential: agrees to use condoms or other methods to ensure effective contraception with partner; agrees not to donate sperm from the inclusion up to 12 weeks after the last dose • QuantiFERON-TB Gold negative; skin testing (e.g., tuberculin skin test) will be allowed if blood testing is not available, or the blood test result is indeterminate • Agrees not to participate in any other interventional study while participating in this protocol	• Pregnancy or lactation • MS relapse in the 6 months prior to dosing • Febrile illness within 4 weeks prior to dosing, or persistent chronic or active infection requiring treatment with systemic antibiotics, antivirals, or antifungals. • Treatment with another investigational drug or other investigational intervention within 3 months prior to baseline • Contraindications for 7-tesla MRI • Presence of any screening laboratory or ECG values outside normal limits that are considered in the PI or MAI's judgment to be clinically significant • Is HIV-positive • Has received any live (attenuated) vaccine (including but not limited to varicella zoster, oral polio, and nasal influenza) within 2 months before dosing • Is receiving strong inducers or inhibitors of CYP3A or CYP2C8 hepatic enzymes • Is receiving anticoagulant/anti-platelet therapies • Has a history or presence of significant other concomitant illness that, according to the PI or MAI's judgment, would adversely affect participation in this study; examples include but are not limited to clinically significant cardiovascular, renal, hepatic, or metabolic disease. • Unwilling to allow coded samples to be processed off-site • Unwilling to have coded samples and/or data saved or used in other studies.

Table C.14: Inclusion and Exclusion Criteria for BRaKE-MS Clinical Trial Non-Treatment Cohort

Inclusion Criteria	**Exclusion Criteria**
• Able to provide informed consent • Willingness to comply with all study procedures and availability for the duration of the study • Male or female, aged ≥ 18 • Diagnosed with multiple sclerosis according to the 2017 revision of the McDonald diagnostic criteria, with no new lesion formation by comparison of baseline MRI scan with a historical MRI scan at least 6 months prior • On intravenous anti-CD20 antibody treatment (specifically intravenous ocrelizumab or rituximab) for at least 6 months, with the most recent dose at most 3 months prior to enrolment. (Participants in this cohort should not be on ofatumumab, another anti-CD20 treatment, at baseline, but they may switch from ocrelizumab or rituximab to ofatumumab after week 48.) • Has a prior 7-tesla MRI scan, no more than 1 year prior to enrolment, demonstrating at least one white matter lesion with a paramagnetic rim • For females of reproductive potential: agrees to use highly effective contraception during study participation • Agrees not to participate in any other interventional study while participating in this protocol	• Pregnancy or lactation • MS relapse in the 6 months prior to baseline • Treatment with another investigational drug or other investigational intervention within 3 months prior to baseline • Contraindications for 7-tesla MRI • Has a history or presence of significant other concomitant illness that, according to the PI or MAI's judgment, would adversely affect participation in this study; examples include but are not limited to clinically significant cardiovascular, renal, hepatic, or metabolic disease. • Unwilling to allow coded samples to be processed off-site • Unwilling to have coded samples and/or data saved or used in other studies

Figure C.2 provides an overview of the PRL+ lesions included in the various analyses based on all PRL ratings. A total of 830 lesions were identified for baseline ratings. After the baseline PRL rating discrepancies between raters were discussed and resolved, 152 PRL+ were initially included in the longitudinal ratings. During longitudinal ratings on the FLASK Application, raters had an option to flag lesions to review or comment on lesions. Based on these comments, 18 lesions originally rated as PRL+ at baseline were excluded from the longitudinal analysis due to reconsideration of original baseline PRL+ rating or image quality. Before removing these 18 lesions, they were all reviewed by both raters together, first in a consensus Flask Application and then in the original NIfTI format. Of the resulting 134 PRL+ lesions, 122 that were visible in all T2* field of view in longitudinal follow-up scans were included in final Longitudinal Rating results (Primary outcome) and 128 PRL+ that were larger than $15mm^3$ were included in the quantitative analysis.

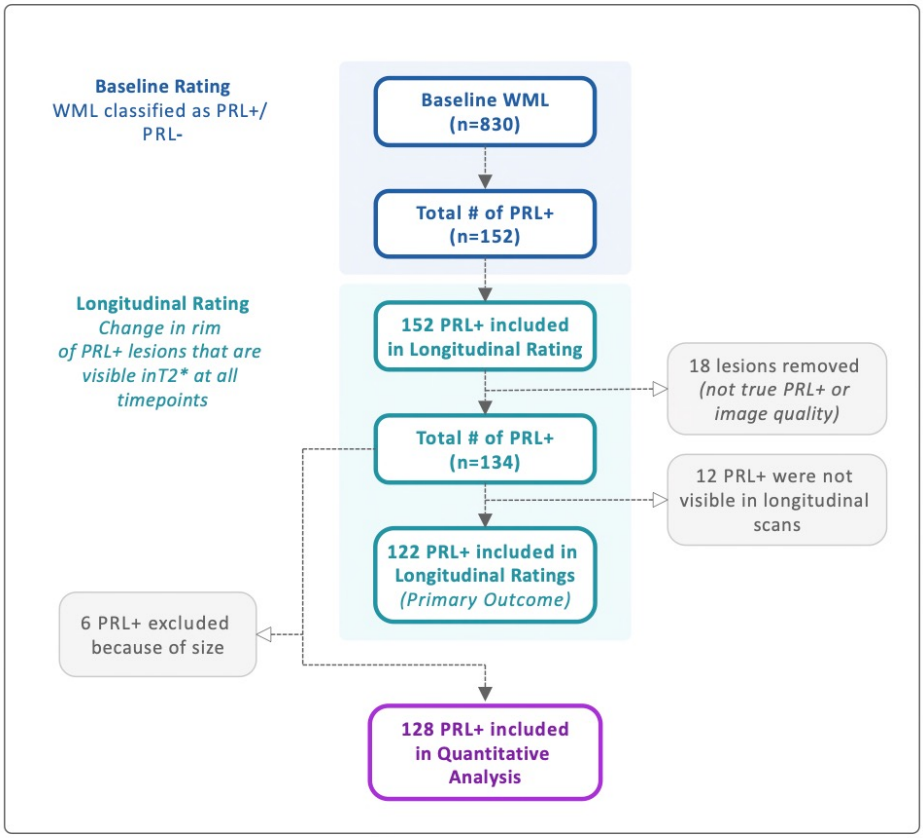


Figure C.2: An overview of PRL+ lesions included in various analyses in the BRaKE-MS clinical trial. In the first radiological baseline ratings of WML, a total of 830 WML were rated as PRL+ or PRL- resulting in 152 PRL+ across all participants. All 152 PRL+ were included in the longitudinal rating FLASK application. At the end of the rating, 18 PRL+ were excluded from the original 152 PRL+ as, upon review raters determined that they were *unratable* (not truly PRL+ or should be excluded due to image quality). This resulted in a final 134 true PRL+ lesions at baseline. Of these, 12 PRL+ were not visible in the follow-up T2*-GRE field of view and therefore excluded from the Primary Outcome Longitudinal ratings. In the quantitative Analysis, we excluded lesions smaller than $15mm^3$, which removed 6/134 PRL+ from this analysis.

C.3.1 Statistical analyses for BRaKE-MS

Table C.15: RCM estimated slopes after removing statistical outliers.

Outcome	**Cohort**	**Comparison**	**Estimate ± SE**	**t-value**	**p-value**
qT1	Non-Treatment	PRL- slope	-0.00040 ± 0.00012	-3.41	0.0009
		PRL+ slope	-0.00022 ± 0.00028	-0.77	0.4439
		PRL slope Δ (PRL+ vs PRL-)	-0.00019 ± 0.00031	-0.60	0.5469
qT1	Treatment	PRL- slope	-0.00022 ± 0.00006	-3.85	0.0001
		PRL+ slope	-0.00027 ± 0.00011	-2.50	0.0129
		PRL slope Δ (PRL+ vs PRL-)	0.00005 ± 0.00012	0.42	0.6749
volume	Non-Treatment	PRL- slope	-0.00076 ± 0.00030	-2.51	0.0135
		PRL+ slope	-0.00058 ± 0.00071	-0.81	0.4221
		PRL slope Δ (PRL+ vs PRL-)	-0.00018 ± 0.00078	-0.24	0.8132
volume	Treatment	PRL- slope	0.00056 ± 0.00015	3.70	0.0002
		PRL+ slope	0.00013 ± 0.00028	0.46	0.6452
		PRL slope Δ (PRL+ vs PRL-)	0.00042 ± 0.00032	1.32	0.1877

Number of statistical outliers removed from each group: normalized lesion qT1 in non-treatment: 9; normalized lesion qT1 in treatment: 22; normalized lesion volume in non-treatment: 11; normalized lesion volume in treatment: 44;

Table C.16: Main (PRL) and simple (PRL+, PRL-) effects for interaction with time in all RMME models.

Cohort	Norm. Dependent Var.	Effect	PRL	F-value	p-value
Non-Treatment	lesion qT1	PRL × Time		0.9	0.45
		PRL × Time	PRL-	10.32	< .0001
		PRL × Time	PRL+	0.44	0.78
Treatment	lesion qT1	PRL × Time		5.77	< .0001
		PRL × Time	PRL-	49.74	< .0001
		PRL × Time	PRL+	5.65	< .0001
Non-Treatment	lesion volume	PRL × Time		1.21	0.30
		PRL × Time	PRL-	8.14	< .0001
		PRL × Time	PRL+	0.25	0.96
Treatment	lesion volume	PRL × Time		1.63	0.13
		PRL × Time	PRL-	8.43	< .0001
		PRL × Time	PRL+	0.22	0.97
Treatment Cohort A	lesion qT1	PRL × Time		4.14	0.0004
		PRL × Time	PRL-	40.51	< .0001
		PRL × Time	PRL+	2.69	0.014
Treatment Cohort B	lesion qT1	PRL × Time		6.20	< .0001
		PRL × Time	PRL-	22.56	< .0001
		PRL × Time	PRL+	5.78	0.0001
Treatment Cohort A	lesion volume	PRL × Time		2.96	0.007
		PRL × Time	PRL-	10.27	< .0001
		PRL × Time	PRL+	0.17	0.99
Treatment Cohort B	lesion volume	PRL × Time		0.14	0.97
		PRL × Time	PRL-	1.99	0.09
		PRL × Time	PRL+	0.25	0.91

Number of statistical outliers removed from each group: normalized lesion qT1 in non-treatment: 9; normalized lesion qT1 in treatment: 22; normalized lesion volume in non-treatment: 11; normalized lesion volume in treatment: 44;

Table C.17: Pairwise Post Hoc Comparisons for median PRL- qT1 in the Non-Treatment Cohort (RRME Model). Results reflect Tukey-Kramer adjusted p-values from repeated measures mixed-effects (RRME) model assessing differences in normalized qT1 between time points, stratified by PRL status.

TP 1	TP 2	Estimate	SE	DF	t-value	Adj. p-value
0	24	0.0377	0.01284	448	2.94	0.1
0	48	0.0645	0.01284	448	5.02	< .0001
0	72	0.07516	0.01284	448	5.85	< .0001
0	96	0.03736	0.01284	448	2.91	0.1
24	48	0.0268	0.01284	448	2.09	0.5
24	72	0.03746	0.01284	448	2.92	0.1
24	96	-0.00033	0.01284	448	-0.03	1
48	72	0.01066	0.01284	448	0.83	0.99
48	96	-0.02714	0.01284	448	-2.11	0.5
72	96	-0.0378	0.01284	448	-2.94	0.1

TP = time point; SE = standard error; DF = degree of freedom

Table C.18: Pairwise Post Hoc Comparisons for qT1 in the Treatment Cohort (RRME Model). Results reflect Tukey-Kramer adjusted p-values from repeated measures mixed-effects (RRME) model assessing differences in normalized volume between time points, stratified by PRL status.

PRL	TP 1	TP 2	Estimate	SE	DF	t-value / Tukey-Kramer p
PRL-	0	24	0.06178	0.006315	2192	9.78 / < .0001
	0	48	0.07318	0.006315	2192	11.59 / < .0001
	0	72	0.08989	0.006315	2192	14.23 / < .0001
	0	96	0.06640	0.006315	2192	10.51 / < .0001
	0	120	0.02506	0.008904	2193	2.81 / 0.2179
	0	144	-0.00450	0.009468	2193	-0.48 / 1
	24	48	0.01141	0.006315	2192	1.81 / 0.8737
	24	72	0.02811	0.006315	2192	4.45 / 0.0007
	24	96	0.00462	0.006315	2192	0.73 / 1
	24	120	-0.03672	0.008904	2193	-4.12 / 0.0031
	24	144	-0.06628	0.009468	2193	-7.00 / < .0001
	48	72	0.01671	0.006315	2192	2.65 / 0.3117
	48	96	-0.00679	0.006315	2192	-1.07 / 0.9986
	48	120	-0.04812	0.008904	2193	-5.40 / < .0001
	48	144	-0.07769	0.009468	2193	-8.21 / < .0001
	72	96	-0.02350	0.006315	2192	-3.72 / 0.0147
	72	120	-0.06483	0.008904	2193	-7.28 / < .0001
	72	144	-0.09439	0.009468	2193	-9.97 / < .0001
	96	120	-0.04134	0.008904	2193	-4.64 / 0.0003
	96	144	-0.07090	0.009468	2193	-7.49 / < .0001
	120	144	-0.02956	0.01077	2192	-2.75 / 0.2537
PRL+	0	24	0.01313	0.01205	2192	1.09 / 0.9984
	0	48	0.00513	0.01205	2192	0.43 / 1
	0	72	0.05700	0.01205	2192	4.73 / 0.0002
	0	96	0.03117	0.01205	2192	2.59 / 0.3499
	0	120	0.02976	0.01593	2194	1.87 / 0.8435
	0	144	-0.01385	0.02014	2195	-0.69 / 1
	24	48	-0.00800	0.01205	2192	-0.66 / 1
	24	72	0.04387	0.01205	2192	3.64 / 0.0196
	24	96	0.01804	0.01205	2192	1.50 / 0.9684
	24	120	0.01663	0.01593	2194	1.04 / 0.9989
	24	144	-0.02698	0.02014	2195	-1.34 / 0.9879
	48	72	0.05187	0.01205	2192	4.30 / 0.0014
	48	96	0.02604	0.01205	2192	2.16 / 0.6586
	48	120	0.02463	0.01593	2194	1.55 / 0.9589
	48	144	-0.01898	0.02014	2195	-0.94 / 0.9996
	72	96	-0.02583	0.01205	2192	-2.14 / 0.6714
	72	120	-0.02724	0.01593	2194	-1.71 / 0.9126
	72	144	-0.07085	0.02014	2195	-3.52 / 0.0298
	96	120	-0.00141	0.01593	2194	-0.09 / 1
	96	144	-0.04502	0.02014	2195	-2.24 / 0.6036
	120	144	-0.04361	0.02187	2193	-1.99 / 0.7715

TP = time point; SE = standard error; DF = degrees of freedom

Table C.19: Pairwise Post Hoc Comparisons for volume in the Treatment Cohort (RRME Model). Results reflect Tukey-Kramer adjusted p-values from repeated measures mixed-effects (RRME) model assessing differences in normalized qT1 between time points, stratified by PRL status.

PRL	TP 1	TP 2	Estimate	SE	DF	t-value / Tukey-Kramer p
PRL-	0	24	-0.06219	0.02011	2176	-3.09 / 0.1087
	0	48	-0.10170	0.02010	2176	-5.06 / < .0001
	0	72	-0.10990	0.02008	2175	-5.48 / < .0001
	0	96	-0.12170	0.02008	2176	-6.06 / < .0001
	0	120	-0.08545	0.02825	2177	-3.03 / 0.13
	0	144	-0.07235	0.03003	2177	-2.41 / 0.4736
	24	48	-0.03946	0.02011	2175	-1.96 / 0.7909
	24	72	-0.04774	0.02010	2176	-2.38 / 0.4988
	24	96	-0.05955	0.02010	2176	-2.96 / 0.1525
	24	120	-0.02326	0.02824	2177	-0.82 / 0.9999
	24	144	-0.01016	0.03002	2177	-0.34 / 1
	48	72	-0.00827	0.02008	2176	-0.41 / 1
	48	96	-0.02009	0.02008	2176	-1.00 / 0.9993
	48	120	0.01620	0.02823	2177	0.57 / 1
	48	144	0.02930	0.03002	2177	0.98 / 0.9995
	72	96	-0.01181	0.02006	2176	-0.59 / 1
	72	120	0.02448	0.02823	2177	0.87 / 0.9999
	72	144	0.03757	0.03001	2177	1.25 / 0.9935
	96	120	0.03629	0.02823	2177	1.29 / 0.9917
	96	144	0.04939	0.03001	2178	1.65 / 0.9338
	120	144	0.01310	0.03411	2176	0.38 / 1

TP = time point; SE = standard error; DF = degrees of freedom

C.3.2 Exploratory analysis of image processing pipeline on lesion qT1 and volume.

Impact of Lesion Segmentation on qT1 Times: Given that the image processing pipeline (Chapter 2.2) allows lesion boundaries to vary across timepoints, temporal changes in lesion segmentation could influence qT1 estimates. Increases in lesion volume could reduce qT1 values by including NAWM voxels, and conversely, decreases in lesion volume could increase qT1 by excluding lower intensity NAWM voxels at the lesion edge. This behaviour was particularly apparent in the PRL- lesions of the BRaKE-MS treatment cohort, where lesion qT1 median (Figure C.3A) appeared to inversely mirror lesion volume (Figure C.3B).

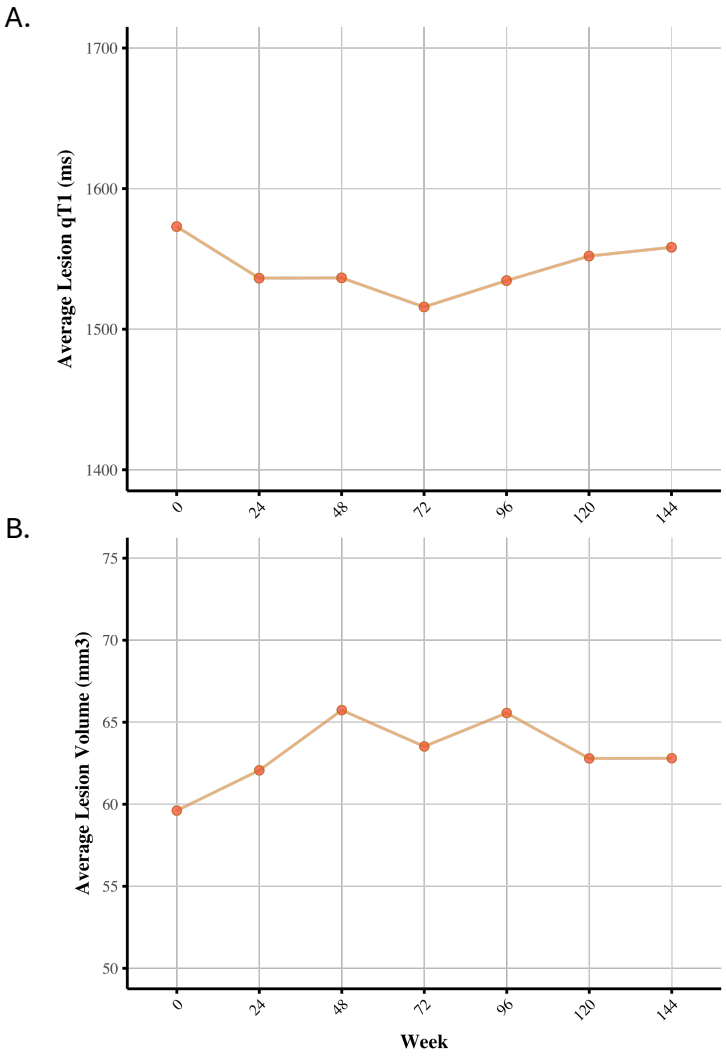


Figure C.3: Inverse relationship between longitudinal lesion qT1 and volume in PRL- lesions. Average (non-confluent) PRL- lesion A) qT1 values and B) lesion volume across timepoints in the BRaKE-MS Treatment Cohort A.

As described in Chapter 2.2, individual lesion segmentations were refined at each timepoint using a boundary correction approach that applies fixed dilation into NAWM, followed by bimodal modelling of qT1 intensities. This method may be less reliable in smaller lesions, making PRL- lesions more vulnerable to segmentation-driven variability. To assess the potential impact of segmentation changes on qT1 values, we conducted an exploratory analysis using a template-based approach in which lesion masks were held constant across timepoints.

Longitudinal skull-stripped qT1 maps were non-linearly registered to a participant-template space using the ANTs multivariate template construction tool. The resulting transformation matrices were applied to each subject's boundary-corrected lesion masks across all time points to align them to the

template space. For each lesion, we generated two composite masks: a union mask, comprising all voxels present in any longitudinal mask, and an intersection mask, including only voxels common to all time points. These composite masks were then transformed back to the native T1 map at each time point. Median qT1 values were extracted using both the union and intersection masks to evaluate whether fixing the lesion mask across time points influenced the observed longitudinal qT1 dynamics.

Figure 2 compares average lesion qT1 values using the original, intersection, and union masks in PRL- lesions from the BRaKE-MS treatment cohort A (Figure C.4A) and for all individual BRaKE-MS participants (Figure C.4B). As expected, qT1 values are highest with the intersection mask as lower-intensity voxels at lesion-NAWM boundaries are excluded. Importantly, the characteristic qT1 trajectory, a decrease followed by an increase, is preserved across all masking strategies. These findings suggest that the qT1 trajectories observed in the PRL- lesions in the BRaKE-MS treatment cohort are unlikely to be driven by segmentation variability.

While average qT1 values are similar between the union and original masks, the intersection mask shows notable differences at specific timepoints (highlighted by arrows in Figure C.4B). In some cases (dark purple arrows), the intersection mask reveals larger fluctuations not present in the original masks, likely due to registration errors in small lesions or those near high-intensity regions like CSF. In other participants, the intersection mask dampens some fluctuations, possibly by reducing partial volume effects included in the original masks in larger lesions. However, both approaches remain susceptible to some random qT1 variations, suggesting that the intersection mask does not offer a consistently better solution.

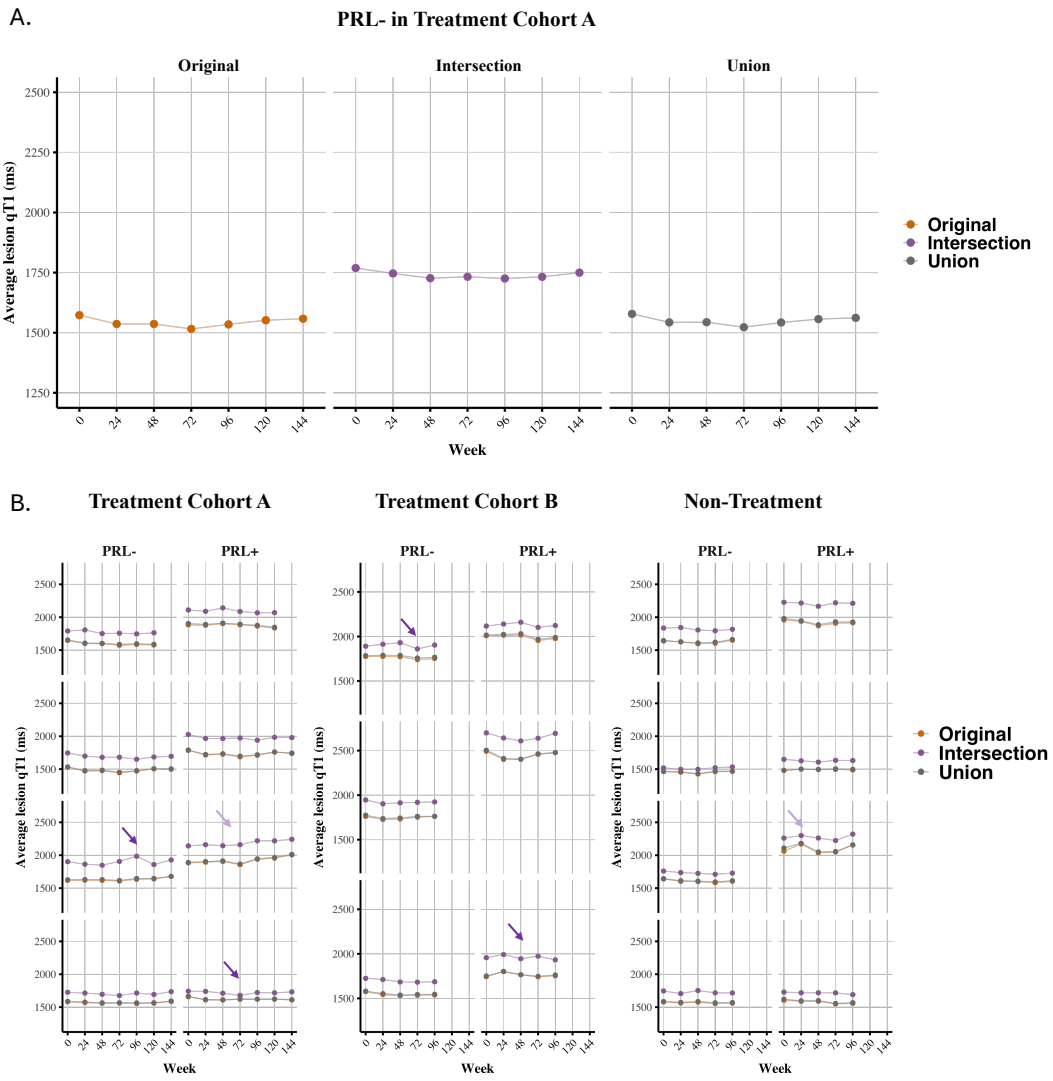


Figure C.4: Comparison of qT1 trajectories using different lesion segmentation masks. A) Average lesion qT1 median in PRL- lesions from the BRaKE-MS treatment cohort A, shown for the original segmentation (orange), intersection mask (purple), and union mask (grey). B) Participant-wise average lesion qT1 medians across the three masking approaches. Dark purple arrows highlight fluctuations introduced by the intersection mask, while light purple arrows indicate timepoint-specific fluctuations reduced in the intersection mask compared to the original segmentation.

Impact of Registration Transformations on Lesion Volume and qT1 Fluctuations:

The boundary-corrected lesion analysis pipeline (described in Chapter 2.2) uses affine registration with 12 degrees of freedom to align baseline lesion segmentations to follow-up time points. Affine transformations capture translation, rotation, scaling, and shear between two images. Non-uniform scaling and shear can artificially alter lesion volumes, expanding or contracting them, even when the underlying anatomy remains unchanged. This effect is especially relevant for small lesions, where minor distortions can lead to noticeable

differences in voxel inclusion. As a result, lesion volume estimates derived from registered masks may reflect not only biological changes, but also geometric transformations introduced by the affine alignment.

To assess whether fluctuations in lesion volume and qT1 were influenced by these registration effects, we extracted the Jacobian determinant (JD) of the affine transformation matrix for each timepoint and participant in the BRaKE-MS trial. The JD is a scalar value that quantifies volume change introduced by the transformation, with values greater than 1 indicating expansion and values less than 1 indicating compression. We then evaluated whether temporal changes in the JD corresponded to the observed changes in lesion volume or qT1, to help determine whether these fluctuations stem from registration-related artifacts.

Figure C.5A shows the average (JD) for each participant, and Figure C.5B shows the participant-wies average percent change in lesion volume from baseline to follow-up. Consistent increase in lesion volume when $JD > 1$ (indicating expansion) and decrease in lesion volume when $JD < 1$ (indicating compression) would suggest that volume changes are at least partially driven by the affine registration scaling. However, this pattern is not observed consistently across participants. Only one participant in the treatment cohort (marked with an asterisk in Figure C.5A and C.5B) shows volume changes that align with JD fluctuations. To further assess this, we normalized lesion volumes by the JD at each timepoint to estimate what lesion size would be in the absence of affine scaling (Figure C.5C). If registration scaling were driving the fluctuations, this correction should reduce variability. However, even in the participant marked with an asterisk in Figure C.5A and C.5B, normalization does not appear to stabilize the lesion volume trajectory.

Finally, figure 3D shows average percent change in lesion qT1 values. While changes in qT1 are not expected to directly mirror JD (e.g., qT1 does not necessarily increase with expansion), registration differences could still influence qT1 measurements. However, visual comparison of Figure C.5A and D does not reveal any consistent or systematic relationship between JD and qT1 changes.

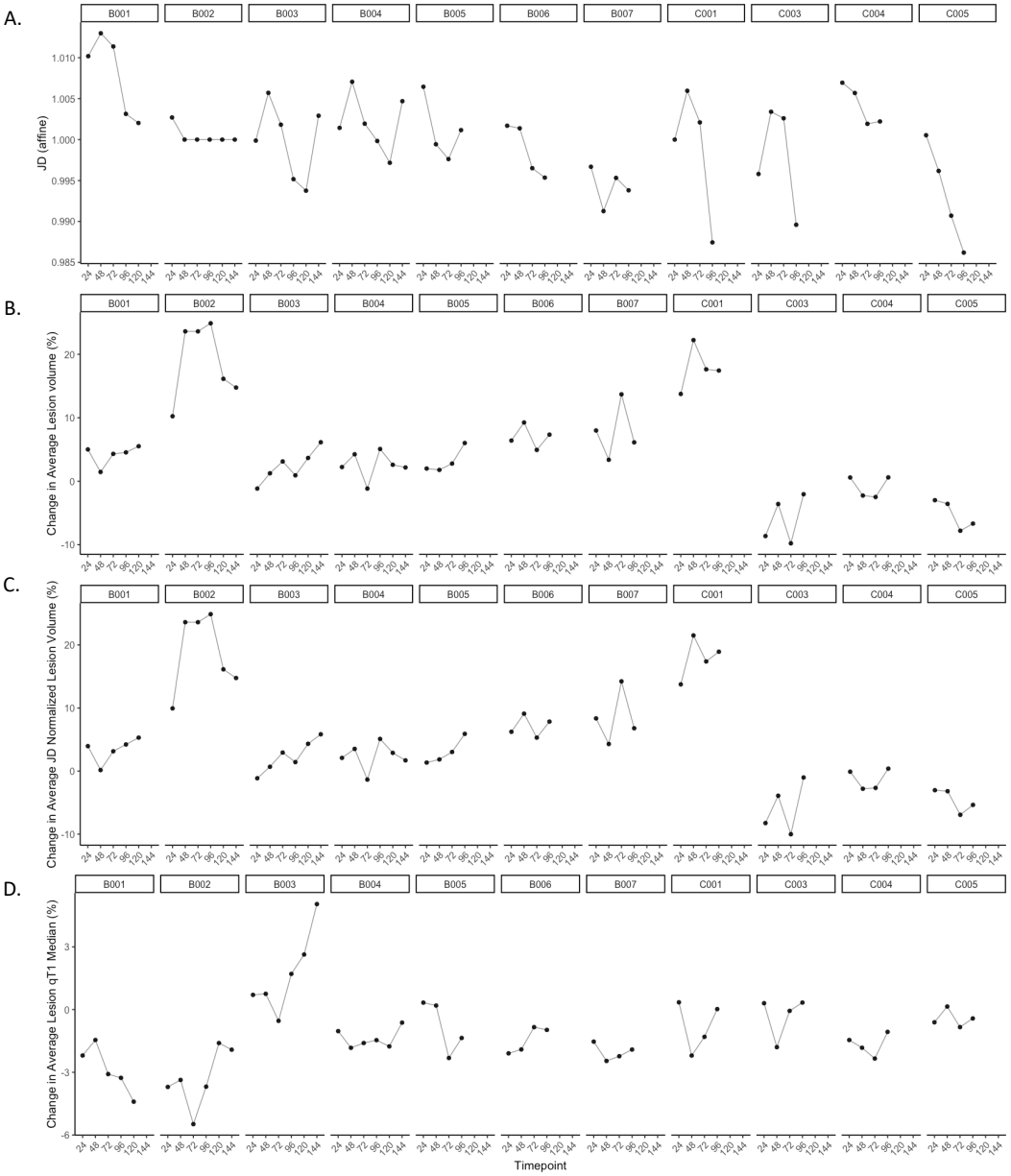


Figure C.5: Comparison of Jacobian determinant (JD) from participant-wise affine registration with changes in lesion volume and qT1 over time. A) Participant-wise JD across follow-up timepoints. B) Average percent change in lesion volume from baseline. C) Average Percent change in lesion volume after normalization of volume by JD, representing estimated volume change without the influence of affine scaling. D) Participant-wise average change in lesion qT1.
An asterisk (*) indicates a participant where registration-related effects may have contributed to observed changes in lesion size.

Appendix D

Supplementary data for Chapter 4.3.4.

D.1 Supplementary data for Chapter 4.1

D.1.1 Pairwise Statistics for Segmentation Overlap Metrics

This appendix provides the detailed pairwise post-hoc statistics underlying the overlap metrics presented in Main Text Figures 4.4 and 4.5.

For all automatic ULF WML segmentation algorithms, overlap metrics were evaluated using the ULF Aided masks as reference. SynthSeg WML masks were down-sampled to the ULF voxel size for these calculations. Pairwise post-hoc test results for the original ULF native space are summarized in Tables D.1–D.7 (corresponding to Main Text Figure 4.4) and D.8–D.14 (corresponding to Main Text Figure 4.5) under the *Original space* column.

To avoid bias, analyses were also performed in SynthSeg's native 1 mm isotropic space, with all other masks up-sampled to 1 mm. Resampling used linear interpolation, with thresholds optimized in 0.05 steps to preserve original lesion volumes. The results of these analyses are shown in Figures D.1 and D.2, with the corresponding pairwise post-hoc statistics included in the same tables (D.1–D.14) under the *1×1×1 mm Space* column.

Table D.1: DSC pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.23 ± 0.02	0.30 ± 0.02	$3.19, 0.039$	0.24 ± 0.02	0.29 ± 0.02	$2.78, 0.154$
MiMoSA	nnU-Net FL	0.23 ± 0.02	0.41 ± 0.03	$5.42, < 0.001$	0.24 ± 0.02	0.41 ± 0.03	$5.18, < 0.001$
MiMoSA	nnU-Net FL/T1w	0.23 ± 0.02	0.41 ± 0.03	$5.84, < 0.001$	0.24 ± 0.02	0.41 ± 0.03	$5.76, < 0.001$
MiMoSA	PLAn 1 FL	0.23 ± 0.02	0.49 ± 0.03	$10.75, < 0.001$	0.24 ± 0.02	0.49 ± 0.03	$10.39, < 0.001$
MiMoSA	PLAn 1 FL/T1w	0.23 ± 0.02	0.49 ± 0.03	$10.31, < 0.001$	0.24 ± 0.02	0.49 ± 0.03	$10.21, < 0.001$
MiMoSA	PLAn 2 FL	0.23 ± 0.02	0.49 ± 0.03	$10.44, < 0.001$	0.24 ± 0.02	0.50 ± 0.03	$10.67, < 0.001$
MiMoSA	PLAn 2 FL/T1w	0.23 ± 0.02	0.48 ± 0.03	$9.55, < 0.001$	0.24 ± 0.02	0.49 ± 0.03	$9.49, < 0.001$
SynthSeg	nnU-Net FL	0.30 ± 0.02	0.41 ± 0.03	$2.23, 0.727$	0.29 ± 0.02	0.41 ± 0.03	$2.40, 0.454$
SynthSeg	nnU-Net FL/T1w	0.30 ± 0.02	0.41 ± 0.03	$2.65, 0.228$	0.29 ± 0.02	0.41 ± 0.03	$2.98, 0.079$
SynthSeg	PLAn 1 FL	0.30 ± 0.02	0.49 ± 0.03	$7.55, < 0.001$	0.29 ± 0.02	0.49 ± 0.03	$7.62, < 0.001$
SynthSeg	PLAn 1 FL/T1w	0.30 ± 0.02	0.49 ± 0.03	$7.12, < 0.001$	0.29 ± 0.02	0.49 ± 0.03	$7.44, < 0.001$
SynthSeg	PLAn 2 FL	0.30 ± 0.02	0.49 ± 0.03	$7.25, < 0.001$	0.29 ± 0.02	0.50 ± 0.03	$7.89, < 0.001$
SynthSeg	PLAn 2 FL/T1w	0.30 ± 0.02	0.48 ± 0.03	$6.36, < 0.001$	0.29 ± 0.02	0.49 ± 0.03	$6.71, < 0.001$
nnU-Net FL	nnU-Net FL/T1w	0.41 ± 0.03	0.41 ± 0.03	$0.42, > 0.99$	0.41 ± 0.03	0.41 ± 0.03	$0.58, > 0.99$
nnU-Net FL	PLAn 1 FL	0.41 ± 0.03	0.49 ± 0.03	$5.33, < 0.001$	0.41 ± 0.03	0.49 ± 0.03	$5.21, < 0.001$
nnU-Net FL	PLAn 1 FL/T1w	0.41 ± 0.03	0.49 ± 0.03	$4.89, < 0.001$	0.41 ± 0.03	0.49 ± 0.03	$5.04, < 0.001$
nnU-Net FL	PLAn 2 FL	0.41 ± 0.03	0.49 ± 0.03	$5.02, < 0.001$	0.41 ± 0.03	0.50 ± 0.03	$5.49, < 0.001$
nnU-Net FL	PLAn 2 FL/T1w	0.41 ± 0.03	0.48 ± 0.03	$4.13, < 0.001$	0.41 ± 0.03	0.49 ± 0.03	$4.31, < 0.001$
nnU-Net FL/T1w	PLAn 1 FL	0.41 ± 0.03	0.49 ± 0.03	$4.91, < 0.001$	0.41 ± 0.03	0.49 ± 0.03	$4.63, < 0.001$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.41 ± 0.03	0.49 ± 0.03	$4.47, < 0.001$	0.41 ± 0.03	0.49 ± 0.03	$4.45, < 0.001$
nnU-Net FL/T1w	PLAn 2 FL	0.41 ± 0.03	0.49 ± 0.03	$4.60, < 0.001$	0.41 ± 0.03	0.50 ± 0.03	$4.91, < 0.001$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.41 ± 0.03	0.48 ± 0.03	$3.71, 0.006$	0.41 ± 0.03	0.49 ± 0.03	$3.73, 0.005$
PLAn 1 FL	PLAn 1 FL/T1w	0.49 ± 0.03	0.49 ± 0.03	$0.44, > 0.99$	0.49 ± 0.03	0.49 ± 0.03	$0.18, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.49 ± 0.03	0.49 ± 0.03	$0.31, > 0.99$	0.49 ± 0.03	0.50 ± 0.03	$0.27, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.49 ± 0.03	0.48 ± 0.03	$1.19, > 0.99$	0.49 ± 0.03	0.49 ± 0.03	$0.90, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.49 ± 0.03	0.49 ± 0.03	$0.13, > 0.99$	0.49 ± 0.03	0.50 ± 0.03	$0.45, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.49 ± 0.03	0.48 ± 0.03	$0.76, > 0.99$	0.49 ± 0.03	0.49 ± 0.03	$0.73, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.49 ± 0.03	0.48 ± 0.03	$0.89, > 0.99$	0.50 ± 0.03	0.49 ± 0.03	$1.18, > 0.99$

Abbreviations: DSC = Dice similarity coefficient; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.2: FPR Metric pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.81 ± 0.02	0.70 ± 0.03	$2.39, 0.474$	0.81 ± 0.02	0.71 ± 0.03	$2.29, 0.614$
MiMoSA	nnU-Net FL	0.81 ± 0.02	0.23 ± 0.03	$13.56, < 0.001$	0.81 ± 0.02	0.23 ± 0.03	$13.67, < 0.001$
MiMoSA	nnU-Net FL/T1w	0.81 ± 0.02	0.22 ± 0.03	$13.94, < 0.001$	0.81 ± 0.02	0.22 ± 0.03	$13.91, < 0.001$
MiMoSA	PLAn 1 FL	0.81 ± 0.02	0.29 ± 0.03	$10.17, < 0.001$	0.81 ± 0.02	0.29 ± 0.03	$10.13, < 0.001$
MiMoSA	PLAn 1 FL/T1w	0.81 ± 0.02	0.30 ± 0.03	$10.46, < 0.001$	0.81 ± 0.02	0.29 ± 0.03	$10.44, < 0.001$
MiMoSA	PLAn 2 FL	0.81 ± 0.02	0.32 ± 0.03	$7.97, < 0.001$	0.81 ± 0.02	0.32 ± 0.03	$7.91, < 0.001$
MiMoSA	PLAn 2 FL/T1w	0.81 ± 0.02	0.32 ± 0.02	$7.88, < 0.001$	0.81 ± 0.02	0.32 ± 0.02	$7.75, < 0.001$
SynthSeg	nnU-Net FL	0.70 ± 0.03	0.23 ± 0.03	$11.17, < 0.001$	0.71 ± 0.03	0.23 ± 0.03	$11.38, < 0.001$
SynthSeg	nnU-Net FL/T1w	0.70 ± 0.03	0.22 ± 0.03	$11.55, < 0.001$	0.71 ± 0.03	0.22 ± 0.03	$11.62, < 0.001$
SynthSeg	PLAn 1 FL	0.70 ± 0.03	0.29 ± 0.03	$7.78, < 0.001$	0.71 ± 0.03	0.29 ± 0.03	$7.84, < 0.001$
SynthSeg	PLAn 1 FL/T1w	0.70 ± 0.03	0.30 ± 0.03	$8.07, < 0.001$	0.71 ± 0.03	0.29 ± 0.03	$8.15, < 0.001$
SynthSeg	PLAn 2 FL	0.70 ± 0.03	0.32 ± 0.03	$5.58, < 0.001$	0.71 ± 0.03	0.32 ± 0.03	$5.62, < 0.001$
SynthSeg	PLAn 2 FL/T1w	0.70 ± 0.03	0.32 ± 0.02	$5.49, < 0.001$	0.71 ± 0.03	0.32 ± 0.02	$5.45, < 0.001$
nnU-Net FL	nnU-Net FL/T1w	0.23 ± 0.03	0.22 ± 0.03	$0.39, > 0.99$	0.23 ± 0.03	0.22 ± 0.03	$0.24, > 0.99$
nnU-Net FL	PLAn 1 FL	0.23 ± 0.03	0.29 ± 0.03	$3.39, 0.020$	0.23 ± 0.03	0.29 ± 0.03	$3.53, 0.011$
nnU-Net FL	PLAn 1 FL/T1w	0.23 ± 0.03	0.30 ± 0.03	$3.10, 0.054$	0.23 ± 0.03	0.29 ± 0.03	$3.23, 0.035$
nnU-Net FL	PLAn 2 FL	0.23 ± 0.03	0.32 ± 0.03	$5.58, < 0.001$	0.23 ± 0.03	0.32 ± 0.03	$5.76, < 0.001$
nnU-Net FL	PLAn 2 FL/T1w	0.23 ± 0.03	0.32 ± 0.02	$5.68, < 0.001$	0.23 ± 0.03	0.32 ± 0.02	$5.92, < 0.001$
nnU-Net FL/T1w	PLAn 1 FL	0.22 ± 0.03	0.29 ± 0.03	$3.78, 0.004$	0.22 ± 0.03	0.29 ± 0.03	$3.78, 0.004$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.22 ± 0.03	0.30 ± 0.03	$3.49, 0.014$	0.22 ± 0.03	0.29 ± 0.03	$3.47, 0.015$
nnU-Net FL/T1w	PLAn 2 FL	0.22 ± 0.03	0.32 ± 0.03	$5.97, < 0.001$	0.22 ± 0.03	0.32 ± 0.03	$6.00, < 0.001$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.22 ± 0.03	0.32 ± 0.02	$6.07, < 0.001$	0.22 ± 0.03	0.32 ± 0.02	$6.16, < 0.001$
PLAn 1 FL	PLAn 1 FL/T1w	0.29 ± 0.03	0.30 ± 0.03	$0.29, > 0.99$	0.29 ± 0.03	0.29 ± 0.03	$0.31, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.29 ± 0.03	0.32 ± 0.03	$2.19, 0.789$	0.29 ± 0.03	0.32 ± 0.03	$2.23, 0.727$
PLAn 1 FL	PLAn 2 FL/T1w	0.29 ± 0.03	0.32 ± 0.02	$2.29, 0.614$	0.29 ± 0.03	0.32 ± 0.02	$2.39, 0.474$
PLAn 1 FL/T1w	PLAn 2 FL	0.30 ± 0.03	0.32 ± 0.03	$2.48, 0.363$	0.29 ± 0.03	0.32 ± 0.03	$2.53, 0.316$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.30 ± 0.03	0.32 ± 0.02	$2.58, 0.275$	0.29 ± 0.03	0.32 ± 0.02	$2.69, 0.197$
PLAn 2 FL	PLAn 2 FL/T1w	0.32 ± 0.03	0.32 ± 0.02	$0.10, > 0.99$	0.32 ± 0.03	0.32 ± 0.02	$0.16, > 0.99$

Abbreviations: FPR = false positive ratio; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.3: FNR pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.57 ± 0.03	0.57 ± 0.02	$1.63, > 0.99$	0.56 ± 0.03	0.59 ± 0.02	$2.15, 0.892$
MiMoSA	nnU-Net FL	0.57 ± 0.03	0.66 ± 0.03	$5.52, < 0.001$	0.56 ± 0.03	0.66 ± 0.03	$5.87, < 0.001$
MiMoSA	nnU-Net FL/T1w	0.57 ± 0.03	0.66 ± 0.03	$5.49, < 0.001$	0.56 ± 0.03	0.66 ± 0.03	$5.71, < 0.001$
MiMoSA	PLAn 1 FL	0.57 ± 0.03	0.56 ± 0.03	$0.39, > 0.99$	0.56 ± 0.03	0.56 ± 0.03	$0.08, > 0.99$
MiMoSA	PLAn 1 FL/T1w	0.57 ± 0.03	0.56 ± 0.03	$0.10, > 0.99$	0.56 ± 0.03	0.56 ± 0.03	$0.10, > 0.99$
MiMoSA	PLAn 2 FL	0.57 ± 0.03	0.54 ± 0.03	$1.68, > 0.99$	0.56 ± 0.03	0.54 ± 0.03	$1.44, > 0.99$
MiMoSA	PLAn 2 FL/T1w	0.57 ± 0.03	0.56 ± 0.03	$0.79, > 0.99$	0.56 ± 0.03	0.55 ± 0.03	$0.44, > 0.99$
SynthSeg	nnU-Net FL	0.57 ± 0.02	0.66 ± 0.03	$3.89, 0.003$	0.59 ± 0.02	0.66 ± 0.03	$3.73, 0.005$
SynthSeg	nnU-Net FL/T1w	0.57 ± 0.02	0.66 ± 0.03	$3.86, 0.003$	0.59 ± 0.02	0.66 ± 0.03	$3.57, 0.010$
SynthSeg	PLAn 1 FL	0.57 ± 0.02	0.56 ± 0.03	$2.02, > 0.99$	0.59 ± 0.02	0.56 ± 0.03	$2.23, 0.727$
SynthSeg	PLAn 1 FL/T1w	0.57 ± 0.02	0.56 ± 0.03	$1.73, > 0.99$	0.59 ± 0.02	0.56 ± 0.03	$2.05, > 0.99$
SynthSeg	PLAn 2 FL	0.57 ± 0.02	0.54 ± 0.03	$3.31, 0.026$	0.59 ± 0.02	0.54 ± 0.03	$3.58, 0.009$
SynthSeg	PLAn 2 FL/T1w	0.57 ± 0.02	0.56 ± 0.03	$2.42, 0.434$	0.59 ± 0.02	0.55 ± 0.03	$2.58, 0.275$
nnU-Net FL	nnU-Net FL/T1w	0.66 ± 0.03	0.66 ± 0.03	$0.03, > 0.99$	0.66 ± 0.03	0.66 ± 0.03	$0.16, > 0.99$
nnU-Net FL	PLAn 1 FL	0.66 ± 0.03	0.56 ± 0.03	$5.91, < 0.001$	0.66 ± 0.03	0.56 ± 0.03	$5.96, < 0.001$
nnU-Net FL	PLAn 1 FL/T1w	0.66 ± 0.03	0.56 ± 0.03	$5.62, < 0.001$	0.66 ± 0.03	0.56 ± 0.03	$5.78, < 0.001$
nnU-Net FL	PLAn 2 FL	0.66 ± 0.03	0.54 ± 0.03	$7.20, < 0.001$	0.66 ± 0.03	0.54 ± 0.03	$7.31, < 0.001$
nnU-Net FL	PLAn 2 FL/T1w	0.66 ± 0.03	0.56 ± 0.03	$6.31, < 0.001$	0.66 ± 0.03	0.55 ± 0.03	$6.31, < 0.001$
nnU-Net FL/T1w	PLAn 1 FL	0.66 ± 0.03	0.56 ± 0.03	$5.87, < 0.001$	0.66 ± 0.03	0.56 ± 0.03	$5.79, < 0.001$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.66 ± 0.03	0.56 ± 0.03	$5.58, < 0.001$	0.66 ± 0.03	0.56 ± 0.03	$5.62, < 0.001$
nnU-Net FL/T1w	PLAn 2 FL	0.66 ± 0.03	0.54 ± 0.03	$7.17, < 0.001$	0.66 ± 0.03	0.54 ± 0.03	$7.15, < 0.001$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.66 ± 0.03	0.56 ± 0.03	$6.28, < 0.001$	0.66 ± 0.03	0.55 ± 0.03	$6.15, < 0.001$
PLAn 1 FL	PLAn 1 FL/T1w	0.56 ± 0.03	0.56 ± 0.03	$0.29, > 0.99$	0.56 ± 0.03	0.56 ± 0.03	$0.18, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.56 ± 0.03	0.54 ± 0.03	$1.29, > 0.99$	0.56 ± 0.03	0.54 ± 0.03	$1.36, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.56 ± 0.03	0.56 ± 0.03	$0.40, > 0.99$	0.56 ± 0.03	0.55 ± 0.03	$0.35, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.56 ± 0.03	0.54 ± 0.03	$1.58, > 0.99$	0.56 ± 0.03	0.54 ± 0.03	$1.53, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.56 ± 0.03	0.56 ± 0.03	$0.69, > 0.99$	0.56 ± 0.03	0.55 ± 0.03	$0.53, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.54 ± 0.03	0.56 ± 0.03	$0.89, > 0.99$	0.54 ± 0.03	0.55 ± 0.03	$1.00, > 0.99$

Abbreviations: FNR = false negative ratio; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.4: FPR cluster pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.90 ± 0.01	0.64 ± 0.02	$2.79, 0.147$	0.90 ± 0.01	0.71 ± 0.02	$2.63, 0.239$
MiMoSA	nnU-Net FL	0.90 ± 0.01	0.14 ± 0.03	$12.28, < 0.001$	0.90 ± 0.01	0.14 ± 0.03	$12.26, < 0.001$
MiMoSA	nnU-Net FL/T1w	0.90 ± 0.01	0.11 ± 0.03	$13.39, < 0.001$	0.90 ± 0.01	0.12 ± 0.03	$13.04, < 0.001$
MiMoSA	PLAn 1 FL	0.90 ± 0.01	0.20 ± 0.03	$10.04, < 0.001$	0.90 ± 0.01	0.20 ± 0.03	$10.20, < 0.001$
MiMoSA	PLAn 1 FL/T1w	0.90 ± 0.01	0.18 ± 0.03	$10.96, < 0.001$	0.90 ± 0.01	0.19 ± 0.03	$10.88, < 0.001$
MiMoSA	PLAn 2 FL	0.90 ± 0.01	0.21 ± 0.03	$9.93, < 0.001$	0.90 ± 0.01	0.22 ± 0.03	$9.55, < 0.001$
MiMoSA	PLAn 2 FL/T1w	0.90 ± 0.01	0.22 ± 0.03	$9.55, < 0.001$	0.90 ± 0.01	0.23 ± 0.03	$9.60, < 0.001$
SynthSeg	nnU-Net FL	0.64 ± 0.02	0.14 ± 0.03	$9.49, < 0.001$	0.71 ± 0.02	0.14 ± 0.03	$9.63, < 0.001$
SynthSeg	nnU-Net FL/T1w	0.64 ± 0.02	0.11 ± 0.03	$10.60, < 0.001$	0.71 ± 0.02	0.12 ± 0.03	$10.41, < 0.001$
SynthSeg	PLAn 1 FL	0.64 ± 0.02	0.20 ± 0.03	$7.25, < 0.001$	0.71 ± 0.02	0.20 ± 0.03	$7.57, < 0.001$
SynthSeg	PLAn 1 FL/T1w	0.64 ± 0.02	0.18 ± 0.03	$8.17, < 0.001$	0.71 ± 0.02	0.19 ± 0.03	$8.25, < 0.001$
SynthSeg	PLAn 2 FL	0.64 ± 0.02	0.21 ± 0.03	$7.13, < 0.001$	0.71 ± 0.02	0.22 ± 0.03	$6.92, < 0.001$
SynthSeg	PLAn 2 FL/T1w	0.64 ± 0.02	0.22 ± 0.03	$6.76, < 0.001$	0.71 ± 0.02	0.23 ± 0.03	$6.97, < 0.001$
nnU-Net FL	nnU-Net FL/T1w	0.14 ± 0.03	0.11 ± 0.03	$1.11, > 0.99$	0.14 ± 0.03	0.12 ± 0.03	$0.77, > 0.99$
nnU-Net FL	PLAn 1 FL	0.14 ± 0.03	0.20 ± 0.03	$2.24, 0.697$	0.14 ± 0.03	0.20 ± 0.03	$2.07, > 0.99$
nnU-Net FL	PLAn 1 FL/T1w	0.14 ± 0.03	0.18 ± 0.03	$1.32, > 0.99$	0.14 ± 0.03	0.19 ± 0.03	$1.39, > 0.99$
nnU-Net FL	PLAn 2 FL	0.14 ± 0.03	0.21 ± 0.03	$2.36, 0.517$	0.14 ± 0.03	0.22 ± 0.03	$2.71, 0.188$
nnU-Net FL	PLAn 2 FL/T1w	0.14 ± 0.03	0.22 ± 0.03	$2.73, 0.179$	0.14 ± 0.03	0.23 ± 0.03	$2.66, 0.217$
nnU-Net FL/T1w	PLAn 1 FL	0.11 ± 0.03	0.20 ± 0.03	$3.36, 0.022$	0.12 ± 0.03	0.20 ± 0.03	$2.84, 0.126$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.11 ± 0.03	0.18 ± 0.03	$2.44, 0.415$	0.12 ± 0.03	0.19 ± 0.03	$2.16, 0.856$
nnU-Net FL/T1w	PLAn 2 FL	0.11 ± 0.03	0.21 ± 0.03	$3.47, 0.015$	0.12 ± 0.03	0.22 ± 0.03	$3.49, 0.014$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.11 ± 0.03	0.22 ± 0.03	$3.84, 0.003$	0.12 ± 0.03	0.23 ± 0.03	$3.44, 0.017$
PLAn 1 FL	PLAn 1 FL/T1w	0.20 ± 0.03	0.18 ± 0.03	$0.92, > 0.99$	0.20 ± 0.03	0.19 ± 0.03	$0.68, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.20 ± 0.03	0.21 ± 0.03	$0.11, > 0.99$	0.20 ± 0.03	0.22 ± 0.03	$0.65, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.20 ± 0.03	0.22 ± 0.03	$0.48, > 0.99$	0.20 ± 0.03	0.23 ± 0.03	$0.60, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.18 ± 0.03	0.21 ± 0.03	$1.03, > 0.99$	0.19 ± 0.03	0.22 ± 0.03	$1.32, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.18 ± 0.03	0.22 ± 0.03	$1.40, > 0.99$	0.19 ± 0.03	0.23 ± 0.03	$1.27, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.21 ± 0.03	0.22 ± 0.03	$0.37, > 0.99$	0.22 ± 0.03	0.23 ± 0.03	$0.05, > 0.99$

Abbreviations: FPR = false positive ration; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.5: FNR cluster pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.35 ± 0.03	0.33 ± 0.02	$1.55, > 0.99$	0.34 ± 0.03	0.29 ± 0.02	$0.50, > 0.99$
MiMoSA	nnU-Net FL	0.35 ± 0.03	0.53 ± 0.03	$7.29, < 0.001$	0.34 ± 0.03	0.53 ± 0.03	$7.44, < 0.001$
MiMoSA	nnU-Net FL/T1w	0.35 ± 0.03	0.55 ± 0.03	$7.83, < 0.001$	0.34 ± 0.03	0.55 ± 0.03	$7.94, < 0.001$
MiMoSA	PLAn 1 FL	0.35 ± 0.03	0.44 ± 0.03	$3.63, 0.008$	0.34 ± 0.03	0.44 ± 0.03	$3.79, 0.004$
MiMoSA	PLAn 1 FL/T1w	0.35 ± 0.03	0.44 ± 0.03	$3.78, 0.004$	0.34 ± 0.03	0.44 ± 0.03	$3.95, 0.002$
MiMoSA	PLAn 2 FL	0.35 ± 0.03	0.44 ± 0.03	$3.63, 0.008$	0.34 ± 0.03	0.44 ± 0.03	$3.65, 0.007$
MiMoSA	PLAn 2 FL/T1w	0.35 ± 0.03	0.45 ± 0.03	$3.92, 0.003$	0.34 ± 0.03	0.45 ± 0.03	$4.10, 0.001$
SynthSeg	nnU-Net FL	0.33 ± 0.02	0.53 ± 0.03	$5.75, < 0.001$	0.29 ± 0.02	0.53 ± 0.03	$6.94, < 0.001$
SynthSeg	nnU-Net FL/T1w	0.33 ± 0.02	0.55 ± 0.03	$6.28, < 0.001$	0.29 ± 0.02	0.55 ± 0.03	$7.44, < 0.001$
SynthSeg	PLAn 1 FL	0.33 ± 0.02	0.44 ± 0.03	$2.08, > 0.99$	0.29 ± 0.02	0.44 ± 0.03	$3.29, 0.028$
SynthSeg	PLAn 1 FL/T1w	0.33 ± 0.02	0.44 ± 0.03	$2.23, 0.727$	0.29 ± 0.02	0.44 ± 0.03	$3.45, 0.015$
SynthSeg	PLAn 2 FL	0.33 ± 0.02	0.44 ± 0.03	$2.08, > 0.99$	0.29 ± 0.02	0.44 ± 0.03	$3.15, 0.046$
SynthSeg	PLAn 2 FL/T1w	0.33 ± 0.02	0.45 ± 0.03	$2.37, 0.495$	0.29 ± 0.02	0.45 ± 0.03	$3.60, 0.009$
nnU-Net FL	nnU-Net FL/T1w	0.53 ± 0.03	0.55 ± 0.03	$0.53, > 0.99$	0.53 ± 0.03	0.55 ± 0.03	$0.50, > 0.99$
nnU-Net FL	PLAn 1 FL	0.53 ± 0.03	0.44 ± 0.03	$3.66, 0.007$	0.53 ± 0.03	0.44 ± 0.03	$3.65, 0.007$
nnU-Net FL	PLAn 1 FL/T1w	0.53 ± 0.03	0.44 ± 0.03	$3.52, 0.012$	0.53 ± 0.03	0.44 ± 0.03	$3.49, 0.014$
nnU-Net FL	PLAn 2 FL	0.53 ± 0.03	0.44 ± 0.03	$3.66, 0.007$	0.53 ± 0.03	0.44 ± 0.03	$3.79, 0.004$
nnU-Net FL	PLAn 2 FL/T1w	0.53 ± 0.03	0.45 ± 0.03	$3.37, 0.021$	0.53 ± 0.03	0.45 ± 0.03	$3.34, 0.023$
nnU-Net FL/T1w	PLAn 1 FL	0.55 ± 0.03	0.44 ± 0.03	$4.20, < 0.001$	0.55 ± 0.03	0.44 ± 0.03	$4.15, < 0.001$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.55 ± 0.03	0.44 ± 0.03	$4.05, 0.001$	0.55 ± 0.03	0.44 ± 0.03	$3.99, 0.002$
nnU-Net FL/T1w	PLAn 2 FL	0.55 ± 0.03	0.44 ± 0.03	$4.20, < 0.001$	0.55 ± 0.03	0.44 ± 0.03	$4.29, < 0.001$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.55 ± 0.03	0.45 ± 0.03	$3.90, 0.003$	0.55 ± 0.03	0.45 ± 0.03	$3.84, 0.003$
PLAn 1 FL	PLAn 1 FL/T1w	0.44 ± 0.03	0.44 ± 0.03	$0.15, > 0.99$	0.44 ± 0.03	0.44 ± 0.03	$0.16, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.44 ± 0.03	0.44 ± 0.03	$0.00, > 0.99$	0.44 ± 0.03	0.44 ± 0.03	$0.15, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.44 ± 0.03	0.45 ± 0.03	$0.29, > 0.99$	0.44 ± 0.03	0.45 ± 0.03	$0.31, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.44 ± 0.03	0.44 ± 0.03	$0.15, > 0.99$	0.44 ± 0.03	0.44 ± 0.03	$0.31, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.44 ± 0.03	0.45 ± 0.03	$0.15, > 0.99$	0.44 ± 0.03	0.45 ± 0.03	$0.15, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.44 ± 0.03	0.45 ± 0.03	$0.29, > 0.99$	0.44 ± 0.03	0.45 ± 0.03	$0.45, > 0.99$

Abbreviations: FNR = false negative ratio; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.6: DER pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	1.26 ± 0.07	0.51 ± 0.05	$7.55, < 0.001$	1.24 ± 0.07	0.31 ± 0.04	$9.78, < 0.001$
MiMoSA	nnU-Net FL	1.26 ± 0.07	0.61 ± 0.07	$6.57, < 0.001$	1.24 ± 0.07	0.60 ± 0.07	$6.02, < 0.001$
MiMoSA	nnU-Net FL/T1w	1.26 ± 0.07	0.62 ± 0.08	$6.84, < 0.001$	1.24 ± 0.07	0.62 ± 0.08	$6.31, < 0.001$
MiMoSA	PLAn 1 FL	1.26 ± 0.07	0.47 ± 0.06	$10.15, < 0.001$	1.24 ± 0.07	0.46 ± 0.06	$9.86, < 0.001$
MiMoSA	PLAn 1 FL/T1w	1.26 ± 0.07	0.47 ± 0.07	$11.09, < 0.001$	1.24 ± 0.07	0.46 ± 0.07	$10.65, < 0.001$
MiMoSA	PLAn 2 FL	1.26 ± 0.07	0.47 ± 0.07	$10.83, < 0.001$	1.24 ± 0.07	0.46 ± 0.07	$10.39, < 0.001$
MiMoSA	PLAn 2 FL/T1w	1.26 ± 0.07	0.47 ± 0.06	$10.75, < 0.001$	1.24 ± 0.07	0.46 ± 0.07	$10.25, < 0.001$
SynthSeg	nnU-Net FL	0.51 ± 0.05	0.61 ± 0.07	$0.98, > 0.99$	0.31 ± 0.04	0.60 ± 0.07	$3.76, 0.005$
SynthSeg	nnU-Net FL/T1w	0.51 ± 0.05	0.62 ± 0.08	$0.71, > 0.99$	0.31 ± 0.04	0.62 ± 0.08	$3.47, 0.015$
SynthSeg	PLAn 1 FL	0.51 ± 0.05	0.47 ± 0.06	$2.60, 0.263$	0.31 ± 0.04	0.46 ± 0.06	$0.08, > 0.99$
SynthSeg	PLAn 1 FL/T1w	0.51 ± 0.05	0.47 ± 0.07	$3.53, 0.011$	0.31 ± 0.04	0.46 ± 0.07	$0.87, > 0.99$
SynthSeg	PLAn 2 FL	0.51 ± 0.05	0.47 ± 0.07	$3.28, 0.029$	0.31 ± 0.04	0.46 ± 0.07	$0.61, > 0.99$
SynthSeg	PLAn 2 FL/T1w	0.51 ± 0.05	0.47 ± 0.06	$3.19, 0.039$	0.31 ± 0.04	0.46 ± 0.07	$0.47, > 0.99$
nnU-Net FL	nnU-Net FL/T1w	0.61 ± 0.07	0.62 ± 0.08	$0.27, > 0.99$	0.60 ± 0.07	0.62 ± 0.08	$0.29, > 0.99$
nnU-Net FL	PLAn 1 FL	0.61 ± 0.07	0.47 ± 0.06	$3.58, 0.009$	0.60 ± 0.07	0.46 ± 0.06	$3.84, 0.003$
nnU-Net FL	PLAn 1 FL/T1w	0.61 ± 0.07	0.47 ± 0.07	$4.52, < 0.001$	0.60 ± 0.07	0.46 ± 0.07	$4.63, < 0.001$
nnU-Net FL	PLAn 2 FL	0.61 ± 0.07	0.47 ± 0.07	$4.26, < 0.001$	0.60 ± 0.07	0.46 ± 0.07	$4.37, < 0.001$
nnU-Net FL	PLAn 2 FL/T1w	0.61 ± 0.07	0.47 ± 0.06	$4.18, < 0.001$	0.60 ± 0.07	0.46 ± 0.07	$4.23, < 0.001$
nnU-Net FL/T1w	PLAn 1 FL	0.62 ± 0.08	0.47 ± 0.06	$3.31, 0.026$	0.62 ± 0.08	0.46 ± 0.06	$3.55, 0.011$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.62 ± 0.08	0.47 ± 0.07	$4.24, < 0.001$	0.62 ± 0.08	0.46 ± 0.07	$4.34, < 0.001$
nnU-Net FL/T1w	PLAn 2 FL	0.62 ± 0.08	0.47 ± 0.07	$3.99, 0.002$	0.62 ± 0.08	0.46 ± 0.07	$4.08, 0.001$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.62 ± 0.08	0.47 ± 0.06	$3.90, 0.003$	0.62 ± 0.08	0.46 ± 0.07	$3.94, 0.002$
PLAn 1 FL	PLAn 1 FL/T1w	0.47 ± 0.06	0.47 ± 0.07	$0.94, > 0.99$	0.46 ± 0.06	0.46 ± 0.07	$0.79, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.47 ± 0.06	0.47 ± 0.07	$0.68, > 0.99$	0.46 ± 0.06	0.46 ± 0.07	$0.53, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.47 ± 0.06	0.47 ± 0.06	$0.60, > 0.99$	0.46 ± 0.06	0.46 ± 0.07	$0.39, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.47 ± 0.07	0.47 ± 0.07	$0.26, > 0.99$	0.46 ± 0.07	0.46 ± 0.07	$0.26, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.47 ± 0.07	0.47 ± 0.06	$0.34, > 0.99$	0.46 ± 0.07	0.46 ± 0.07	$0.40, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.47 ± 0.07	0.47 ± 0.06	$0.08, > 0.99$	0.46 ± 0.07	0.46 ± 0.07	$0.15, > 0.99$

Abbreviations: DER = detection error rate; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.7: OER pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.28 ± 0.02	0.90 ± 0.03	$13.52, < 0.001$	0.29 ± 0.02	1.10 ± 0.04	$14.27, < 0.001$
MiMoSA	nnU-Net FL	0.28 ± 0.02	0.58 ± 0.04	$7.78, < 0.001$	0.29 ± 0.02	0.58 ± 0.04	$7.63, < 0.001$
MiMoSA	nnU-Net FL/T1w	0.28 ± 0.02	0.56 ± 0.04	$6.83, < 0.001$	0.29 ± 0.02	0.56 ± 0.04	$6.47, < 0.001$
MiMoSA	PLAn 1 FL	0.28 ± 0.02	0.56 ± 0.03	$5.34, < 0.001$	0.29 ± 0.02	0.56 ± 0.03	$4.68, < 0.001$
MiMoSA	PLAn 1 FL/T1w	0.28 ± 0.02	0.55 ± 0.03	$6.10, < 0.001$	0.29 ± 0.02	0.55 ± 0.03	$5.44, < 0.001$
MiMoSA	PLAn 2 FL	0.28 ± 0.02	0.54 ± 0.03	$5.70, < 0.001$	0.29 ± 0.02	0.54 ± 0.03	$5.08, < 0.001$
MiMoSA	PLAn 2 FL/T1w	0.28 ± 0.02	0.57 ± 0.03	$6.63, < 0.001$	0.29 ± 0.02	0.57 ± 0.03	$6.00, < 0.001$
SynthSeg	nnU-Net FL	0.90 ± 0.03	0.58 ± 0.04	$5.75, < 0.001$	1.10 ± 0.04	0.58 ± 0.04	$6.63, < 0.001$
SynthSeg	nnU-Net FL/T1w	0.90 ± 0.03	0.56 ± 0.04	$6.70, < 0.001$	1.10 ± 0.04	0.56 ± 0.04	$7.79, < 0.001$
SynthSeg	PLAn 1 FL	0.90 ± 0.03	0.56 ± 0.03	$8.18, < 0.001$	1.10 ± 0.04	0.56 ± 0.03	$9.59, < 0.001$
SynthSeg	PLAn 1 FL/T1w	0.90 ± 0.03	0.55 ± 0.03	$7.42, < 0.001$	1.10 ± 0.04	0.55 ± 0.03	$8.83, < 0.001$
SynthSeg	PLAn 2 FL	0.90 ± 0.03	0.54 ± 0.03	$7.83, < 0.001$	1.10 ± 0.04	0.54 ± 0.03	$9.18, < 0.001$
SynthSeg	PLAn 2 FL/T1w	0.90 ± 0.03	0.57 ± 0.03	$6.89, < 0.001$	1.10 ± 0.04	0.57 ± 0.03	$8.26, < 0.001$
nnU-Net FL	nnU-Net FL/T1w	0.58 ± 0.04	0.56 ± 0.04	$0.95, > 0.99$	0.58 ± 0.04	0.56 ± 0.04	$1.16, > 0.99$
nnU-Net FL	PLAn 1 FL	0.58 ± 0.04	0.56 ± 0.03	$2.44, 0.415$	0.58 ± 0.04	0.56 ± 0.03	$2.95, 0.088$
nnU-Net FL	PLAn 1 FL/T1w	0.58 ± 0.04	0.55 ± 0.03	$1.68, > 0.99$	0.58 ± 0.04	0.55 ± 0.03	$2.19, 0.789$
nnU-Net FL	PLAn 2 FL	0.58 ± 0.04	0.54 ± 0.03	$2.08, > 0.99$	0.58 ± 0.04	0.54 ± 0.03	$2.55, 0.302$
nnU-Net FL	PLAn 2 FL/T1w	0.58 ± 0.04	0.57 ± 0.03	$1.15, > 0.99$	0.58 ± 0.04	0.57 ± 0.03	$1.63, > 0.99$
nnU-Net FL/T1w	PLAn 1 FL	0.56 ± 0.04	0.56 ± 0.03	$1.49, > 0.99$	0.56 ± 0.04	0.56 ± 0.03	$1.79, > 0.99$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.56 ± 0.04	0.55 ± 0.03	$0.73, > 0.99$	0.56 ± 0.04	0.55 ± 0.03	$1.03, > 0.99$
nnU-Net FL/T1w	PLAn 2 FL	0.56 ± 0.04	0.54 ± 0.03	$1.13, > 0.99$	0.56 ± 0.04	0.54 ± 0.03	$1.39, > 0.99$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.56 ± 0.04	0.57 ± 0.03	$0.19, > 0.99$	0.56 ± 0.04	0.57 ± 0.03	$0.47, > 0.99$
PLAn 1 FL	PLAn 1 FL/T1w	0.56 ± 0.03	0.55 ± 0.03	$0.76, > 0.99$	0.56 ± 0.03	0.55 ± 0.03	$0.76, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.56 ± 0.03	0.54 ± 0.03	$0.35, > 0.99$	0.56 ± 0.03	0.54 ± 0.03	$0.40, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.56 ± 0.03	0.57 ± 0.03	$1.29, > 0.99$	0.56 ± 0.03	0.57 ± 0.03	$1.32, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.55 ± 0.03	0.54 ± 0.03	$0.40, > 0.99$	0.55 ± 0.03	0.54 ± 0.03	$0.35, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.55 ± 0.03	0.57 ± 0.03	$0.53, > 0.99$	0.55 ± 0.03	0.57 ± 0.03	$0.56, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.54 ± 0.03	0.57 ± 0.03	$0.94, > 0.99$	0.54 ± 0.03	0.57 ± 0.03	$0.92, > 0.99$

Abbreviations: OER = outline error rate; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.8: DSC pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference in ULF Blinded Subset. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.20 ± 0.03	0.29 ± 0.03	$3.19, 0.039$	0.21 ± 0.04	0.28 ± 0.03	$2.36, 0.658$
MiMoSA	nnU-Net FL	0.20 ± 0.03	0.37 ± 0.05	$5.42, < 0.001$	0.21 ± 0.04	0.38 ± 0.05	$3.53, 0.015$
MiMoSA	nnU-Net FL/T1w	0.20 ± 0.03	0.37 ± 0.05	$5.84, < 0.001$	0.21 ± 0.04	0.37 ± 0.05	$3.42, 0.023$
MiMoSA	PLAn 1 FL	0.20 ± 0.03	0.43 ± 0.05	$10.75, < 0.001$	0.21 ± 0.04	0.43 ± 0.05	$5.96, < 0.001$
MiMoSA	PLAn 1 FL/T1w	0.20 ± 0.03	0.45 ± 0.05	$10.31, < 0.001$	0.21 ± 0.04	0.45 ± 0.05	$6.79, < 0.001$
MiMoSA	PLAn 2 FL	0.20 ± 0.03	0.44 ± 0.05	$10.44, < 0.001$	0.21 ± 0.04	0.44 ± 0.05	$6.25, < 0.001$
MiMoSA	PLAn 2 FL/T1w	0.20 ± 0.03	0.45 ± 0.05	$9.55, < 0.001$	0.21 ± 0.04	0.45 ± 0.05	$6.45, < 0.001$
SynthSeg	nnU-Net FL	0.29 ± 0.03	0.37 ± 0.05	$2.23, 0.727$	0.28 ± 0.03	0.38 ± 0.05	$1.17, > 0.99$
SynthSeg	nnU-Net FL/T1w	0.29 ± 0.03	0.37 ± 0.05	$2.65, 0.228$	0.28 ± 0.03	0.37 ± 0.05	$1.06, > 0.99$
SynthSeg	PLAn 1 FL	0.29 ± 0.03	0.43 ± 0.05	$7.55, < 0.001$	0.28 ± 0.03	0.43 ± 0.05	$3.60, 0.012$
SynthSeg	PLAn 1 FL/T1w	0.29 ± 0.03	0.45 ± 0.05	$7.12, < 0.001$	0.28 ± 0.03	0.45 ± 0.05	$4.43, < 0.001$
SynthSeg	PLAn 2 FL	0.29 ± 0.03	0.44 ± 0.05	$7.25, < 0.001$	0.28 ± 0.03	0.44 ± 0.05	$3.89, 0.004$
SynthSeg	PLAn 2 FL/T1w	0.29 ± 0.03	0.45 ± 0.05	$6.36, < 0.001$	0.28 ± 0.03	0.45 ± 0.05	$4.09, 0.002$
nnU-Net FL	nnU-Net FL/T1w	0.37 ± 0.05	0.37 ± 0.05	$0.42, > 0.99$	0.38 ± 0.05	0.37 ± 0.05	$0.11, > 0.99$
nnU-Net FL	PLAn 1 FL	0.37 ± 0.05	0.43 ± 0.05	$5.33, < 0.001$	0.38 ± 0.05	0.43 ± 0.05	$2.43, 0.548$
nnU-Net FL	PLAn 1 FL/T1w	0.37 ± 0.05	0.45 ± 0.05	$4.89, < 0.001$	0.38 ± 0.05	0.45 ± 0.05	$3.26, 0.040$
nnU-Net FL	PLAn 2 FL	0.37 ± 0.05	0.44 ± 0.05	$5.02, < 0.001$	0.38 ± 0.05	0.44 ± 0.05	$2.72, 0.235$
nnU-Net FL	PLAn 2 FL/T1w	0.37 ± 0.05	0.45 ± 0.05	$4.13, < 0.001$	0.38 ± 0.05	0.45 ± 0.05	$2.92, 0.125$
nnU-Net FL/T1w	PLAn 1 FL	0.37 ± 0.05	0.43 ± 0.05	$4.91, < 0.001$	0.37 ± 0.05	0.43 ± 0.05	$2.54, 0.400$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.37 ± 0.05	0.45 ± 0.05	$4.47, < 0.001$	0.37 ± 0.05	0.45 ± 0.05	$3.37, 0.027$
nnU-Net FL/T1w	PLAn 2 FL	0.37 ± 0.05	0.44 ± 0.05	$4.60, < 0.001$	0.37 ± 0.05	0.44 ± 0.05	$2.83, 0.167$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.37 ± 0.05	0.45 ± 0.05	$3.71, 0.006$	0.37 ± 0.05	0.45 ± 0.05	$3.03, 0.087$
PLAn 1 FL	PLAn 1 FL/T1w	0.43 ± 0.05	0.45 ± 0.05	$0.44, > 0.99$	0.43 ± 0.05	0.45 ± 0.05	$0.83, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.43 ± 0.05	0.44 ± 0.05	$0.31, > 0.99$	0.43 ± 0.05	0.44 ± 0.05	$0.29, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.43 ± 0.05	0.45 ± 0.05	$1.19, > 0.99$	0.43 ± 0.05	0.45 ± 0.05	$0.49, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.45 ± 0.05	0.44 ± 0.05	$0.13, > 0.99$	0.45 ± 0.05	0.44 ± 0.05	$0.54, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.45 ± 0.05	0.45 ± 0.05	$0.76, > 0.99$	0.45 ± 0.05	0.45 ± 0.05	$0.34, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.44 ± 0.05	0.45 ± 0.05	$0.89, > 0.99$	0.44 ± 0.05	0.45 ± 0.05	$0.20, > 0.99$

Abbreviations: DSC = Dice similarity coefficient; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.9: FPR pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference in ULF Blinded Subset. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.84 ± 0.03	0.71 ± 0.04	$2.39, 0.474$	0.84 ± 0.03	0.72 ± 0.04	$1.35, > 0.99$
MiMoSA	nnU-Net FL	0.84 ± 0.03	0.21 ± 0.05	$13.56, < 0.001$	0.84 ± 0.03	0.21 ± 0.04	$8.38, < 0.001$
MiMoSA	nnU-Net FL/T1w	0.84 ± 0.03	0.21 ± 0.04	$13.94, < 0.001$	0.84 ± 0.03	0.21 ± 0.04	$8.72, < 0.001$
MiMoSA	PLAn 1 FL	0.84 ± 0.03	0.26 ± 0.04	$10.17, < 0.001$	0.84 ± 0.03	0.27 ± 0.04	$6.56, < 0.001$
MiMoSA	PLAn 1 FL/T1w	0.84 ± 0.03	0.28 ± 0.04	$10.46, < 0.001$	0.84 ± 0.03	0.28 ± 0.04	$6.85, < 0.001$
MiMoSA	PLAn 2 FL	0.84 ± 0.03	0.29 ± 0.04	$7.97, < 0.001$	0.84 ± 0.03	0.30 ± 0.04	$5.37, < 0.001$
MiMoSA	PLAn 2 FL/T1w	0.84 ± 0.03	0.30 ± 0.04	$7.88, < 0.001$	0.84 ± 0.03	0.30 ± 0.04	$5.60, < 0.001$
SynthSeg	nnU-Net FL	0.71 ± 0.04	0.21 ± 0.05	$11.17, < 0.001$	0.72 ± 0.04	0.21 ± 0.04	$7.03, < 0.001$
SynthSeg	nnU-Net FL/T1w	0.71 ± 0.04	0.21 ± 0.04	$11.55, < 0.001$	0.72 ± 0.04	0.21 ± 0.04	$7.37, < 0.001$
SynthSeg	PLAn 1 FL	0.71 ± 0.04	0.26 ± 0.04	$7.78, < 0.001$	0.72 ± 0.04	0.27 ± 0.04	$5.21, < 0.001$
SynthSeg	PLAn 1 FL/T1w	0.71 ± 0.04	0.28 ± 0.04	$8.07, < 0.001$	0.72 ± 0.04	0.28 ± 0.04	$5.51, < 0.001$
SynthSeg	PLAn 2 FL	0.71 ± 0.04	0.29 ± 0.04	$5.58, < 0.001$	0.72 ± 0.04	0.30 ± 0.04	$4.02, 0.002$
SynthSeg	PLAn 2 FL/T1w	0.71 ± 0.04	0.30 ± 0.04	$5.49, < 0.001$	0.72 ± 0.04	0.30 ± 0.04	$4.25, < 0.001$
nnU-Net FL	nnU-Net FL/T1w	0.21 ± 0.05	0.21 ± 0.04	$0.39, > 0.99$	0.21 ± 0.04	0.21 ± 0.04	$0.34, > 0.99$
nnU-Net FL	PLAn 1 FL	0.21 ± 0.05	0.26 ± 0.04	$3.39, 0.020$	0.21 ± 0.04	0.27 ± 0.04	$1.82, > 0.99$
nnU-Net FL	PLAn 1 FL/T1w	0.21 ± 0.05	0.28 ± 0.04	$3.10, 0.054$	0.21 ± 0.04	0.28 ± 0.04	$1.53, > 0.99$
nnU-Net FL	PLAn 2 FL	0.21 ± 0.05	0.29 ± 0.04	$5.58, < 0.001$	0.21 ± 0.04	0.30 ± 0.04	$3.01, 0.094$
nnU-Net FL	PLAn 2 FL/T1w	0.21 ± 0.05	0.30 ± 0.04	$5.68, < 0.001$	0.21 ± 0.04	0.30 ± 0.04	$2.79, 0.192$
nnU-Net FL/T1w	PLAn 1 FL	0.21 ± 0.04	0.26 ± 0.04	$3.78, 0.004$	0.21 ± 0.04	0.27 ± 0.04	$2.16, > 0.99$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.21 ± 0.04	0.28 ± 0.04	$3.49, 0.014$	0.21 ± 0.04	0.28 ± 0.04	$1.86, > 0.99$
nnU-Net FL/T1w	PLAn 2 FL	0.21 ± 0.04	0.29 ± 0.04	$5.97, < 0.001$	0.21 ± 0.04	0.30 ± 0.04	$3.35, 0.029$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.21 ± 0.04	0.30 ± 0.04	$6.07, < 0.001$	0.21 ± 0.04	0.30 ± 0.04	$3.12, 0.064$
PLAn 1 FL	PLAn 1 FL/T1w	0.26 ± 0.04	0.28 ± 0.04	$0.29, > 0.99$	0.27 ± 0.04	0.28 ± 0.04	$0.29, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.26 ± 0.04	0.29 ± 0.04	$2.19, 0.789$	0.27 ± 0.04	0.30 ± 0.04	$1.19, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.26 ± 0.04	0.30 ± 0.04	$2.29, 0.614$	0.27 ± 0.04	0.30 ± 0.04	$0.97, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.28 ± 0.04	0.29 ± 0.04	$2.48, 0.363$	0.28 ± 0.04	0.30 ± 0.04	$1.48, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.28 ± 0.04	0.30 ± 0.04	$2.58, 0.275$	0.28 ± 0.04	0.30 ± 0.04	$1.26, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.29 ± 0.04	0.30 ± 0.04	$0.10, > 0.99$	0.30 ± 0.04	0.30 ± 0.04	$0.22, > 0.99$

Abbreviations: FPR = false positive ratio; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.10: FNR pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference in ULF Blinded Subset. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.58 ± 0.05	0.55 ± 0.03	$1.63, > 0.99$	0.58 ± 0.05	0.56 ± 0.03	$2.04, > 0.99$
MiMoSA	nnU-Net FL	0.58 ± 0.05	0.68 ± 0.04	$5.52, < 0.001$	0.58 ± 0.05	0.68 ± 0.04	$4.74, < 0.001$
MiMoSA	nnU-Net FL/T1w	0.58 ± 0.05	0.69 ± 0.05	$5.49, < 0.001$	0.58 ± 0.05	0.69 ± 0.05	$5.10, < 0.001$
MiMoSA	PLAn 1 FL	0.58 ± 0.05	0.63 ± 0.05	$0.39, > 0.99$	0.58 ± 0.05	0.63 ± 0.05	$2.56, 0.375$
MiMoSA	PLAn 1 FL/T1w	0.58 ± 0.05	0.60 ± 0.04	$0.10, > 0.99$	0.58 ± 0.05	0.60 ± 0.05	$1.46, > 0.99$
MiMoSA	PLAn 2 FL	0.58 ± 0.05	0.60 ± 0.05	$1.68, > 0.99$	0.58 ± 0.05	0.60 ± 0.05	$1.39, > 0.99$
MiMoSA	PLAn 2 FL/T1w	0.58 ± 0.05	0.60 ± 0.05	$0.79, > 0.99$	0.58 ± 0.05	0.59 ± 0.05	$1.21, > 0.99$
SynthSeg	nnU-Net FL	0.55 ± 0.03	0.68 ± 0.04	$3.89, 0.003$	0.56 ± 0.03	0.68 ± 0.04	$2.70, 0.252$
SynthSeg	nnU-Net FL/T1w	0.55 ± 0.03	0.69 ± 0.05	$3.86, 0.003$	0.56 ± 0.03	0.69 ± 0.05	$3.06, 0.081$
SynthSeg	PLAn 1 FL	0.55 ± 0.03	0.63 ± 0.05	$2.02, > 0.99$	0.56 ± 0.03	0.63 ± 0.05	$0.52, > 0.99$
SynthSeg	PLAn 1 FL/T1w	0.55 ± 0.03	0.60 ± 0.04	$1.73, > 0.99$	0.56 ± 0.03	0.60 ± 0.05	$0.58, > 0.99$
SynthSeg	PLAn 2 FL	0.55 ± 0.03	0.60 ± 0.05	$3.31, 0.026$	0.56 ± 0.03	0.60 ± 0.05	$0.65, > 0.99$
SynthSeg	PLAn 2 FL/T1w	0.55 ± 0.03	0.60 ± 0.05	$2.42, 0.434$	0.56 ± 0.03	0.59 ± 0.05	$0.83, > 0.99$
nnU-Net FL	nnU-Net FL/T1w	0.68 ± 0.04	0.69 ± 0.05	$0.03, > 0.99$	0.68 ± 0.04	0.69 ± 0.05	$0.36, > 0.99$
nnU-Net FL	PLAn 1 FL	0.68 ± 0.04	0.63 ± 0.05	$5.91, < 0.001$	0.68 ± 0.04	0.63 ± 0.05	$2.18, > 0.99$
nnU-Net FL	PLAn 1 FL/T1w	0.68 ± 0.04	0.60 ± 0.04	$5.62, < 0.001$	0.68 ± 0.04	0.60 ± 0.05	$3.28, 0.037$
nnU-Net FL	PLAn 2 FL	0.68 ± 0.04	0.60 ± 0.05	$7.20, < 0.001$	0.68 ± 0.04	0.60 ± 0.05	$3.35, 0.029$
nnU-Net FL	PLAn 2 FL/T1w	0.68 ± 0.04	0.60 ± 0.05	$6.31, < 0.001$	0.68 ± 0.04	0.59 ± 0.05	$3.53, 0.015$
nnU-Net FL/T1w	PLAn 1 FL	0.69 ± 0.05	0.63 ± 0.05	$5.87, < 0.001$	0.69 ± 0.05	0.63 ± 0.05	$2.54, 0.400$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.69 ± 0.05	0.60 ± 0.04	$5.58, < 0.001$	0.69 ± 0.05	0.60 ± 0.05	$3.64, 0.010$
nnU-Net FL/T1w	PLAn 2 FL	0.69 ± 0.05	0.60 ± 0.05	$7.17, < 0.001$	0.69 ± 0.05	0.60 ± 0.05	$3.71, 0.007$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.69 ± 0.05	0.60 ± 0.05	$6.28, < 0.001$	0.69 ± 0.05	0.59 ± 0.05	$3.89, 0.004$
PLAn 1 FL	PLAn 1 FL/T1w	0.63 ± 0.05	0.60 ± 0.04	$0.29, > 0.99$	0.63 ± 0.05	0.60 ± 0.05	$1.10, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.63 ± 0.05	0.60 ± 0.05	$1.29, > 0.99$	0.63 ± 0.05	0.60 ± 0.05	$1.17, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.63 ± 0.05	0.60 ± 0.05	$0.40, > 0.99$	0.63 ± 0.05	0.59 ± 0.05	$1.35, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.60 ± 0.04	0.60 ± 0.05	$1.58, > 0.99$	0.60 ± 0.05	0.60 ± 0.05	$0.07, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.60 ± 0.04	0.60 ± 0.05	$0.69, > 0.99$	0.60 ± 0.05	0.59 ± 0.05	$0.25, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.60 ± 0.05	0.60 ± 0.05	$0.89, > 0.99$	0.60 ± 0.05	0.59 ± 0.05	$0.18, > 0.99$

Abbreviations: FNR = false negative ratio; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.11: FPR Cluster pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference in ULF Blinded Subset. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.93 ± 0.01	0.67 ± 0.03	$2.79, 0.147$	0.93 ± 0.01	0.73 ± 0.03	$1.57, > 0.99$
MiMoSA	nnU-Net FL	0.93 ± 0.01	0.13 ± 0.04	$12.28, < 0.001$	0.93 ± 0.01	0.13 ± 0.04	$7.98, < 0.001$
MiMoSA	nnU-Net FL/T1w	0.93 ± 0.01	0.10 ± 0.04	$13.39, < 0.001$	0.93 ± 0.01	0.10 ± 0.04	$8.49, < 0.001$
MiMoSA	PLAn 1 FL	0.93 ± 0.01	0.20 ± 0.05	$10.04, < 0.001$	0.93 ± 0.01	0.20 ± 0.05	$6.20, < 0.001$
MiMoSA	PLAn 1 FL/T1w	0.93 ± 0.01	0.18 ± 0.05	$10.96, < 0.001$	0.93 ± 0.01	0.18 ± 0.05	$7.44, < 0.001$
MiMoSA	PLAn 2 FL	0.93 ± 0.01	0.21 ± 0.04	$9.93, < 0.001$	0.93 ± 0.01	0.22 ± 0.04	$6.25, < 0.001$
MiMoSA	PLAn 2 FL/T1w	0.93 ± 0.01	0.20 ± 0.05	$9.55, < 0.001$	0.93 ± 0.01	0.21 ± 0.05	$6.56, < 0.001$
SynthSeg	nnU-Net FL	0.67 ± 0.03	0.13 ± 0.04	$9.49, < 0.001$	0.73 ± 0.03	0.13 ± 0.04	$6.41, < 0.001$
SynthSeg	nnU-Net FL/T1w	0.67 ± 0.03	0.10 ± 0.04	$10.60, < 0.001$	0.73 ± 0.03	0.10 ± 0.04	$6.92, < 0.001$
SynthSeg	PLAn 1 FL	0.67 ± 0.03	0.20 ± 0.05	$7.25, < 0.001$	0.73 ± 0.03	0.20 ± 0.05	$4.63, < 0.001$
SynthSeg	PLAn 1 FL/T1w	0.67 ± 0.03	0.18 ± 0.05	$8.17, < 0.001$	0.73 ± 0.03	0.18 ± 0.05	$5.87, < 0.001$
SynthSeg	PLAn 2 FL	0.67 ± 0.03	0.21 ± 0.04	$7.13, < 0.001$	0.73 ± 0.03	0.22 ± 0.04	$4.67, < 0.001$
SynthSeg	PLAn 2 FL/T1w	0.67 ± 0.03	0.20 ± 0.05	$6.76, < 0.001$	0.73 ± 0.03	0.21 ± 0.05	$4.99, < 0.001$
nnU-Net FL	nnU-Net FL/T1w	0.13 ± 0.04	0.10 ± 0.04	$1.11, > 0.99$	0.13 ± 0.04	0.10 ± 0.04	$0.52, > 0.99$
nnU-Net FL	PLAn 1 FL	0.13 ± 0.04	0.20 ± 0.05	$2.24, 0.697$	0.13 ± 0.04	0.20 ± 0.05	$1.77, > 0.99$
nnU-Net FL	PLAn 1 FL/T1w	0.13 ± 0.04	0.18 ± 0.05	$1.32, > 0.99$	0.13 ± 0.04	0.18 ± 0.05	$0.54, > 0.99$
nnU-Net FL	PLAn 2 FL	0.13 ± 0.04	0.21 ± 0.04	$2.36, 0.517$	0.13 ± 0.04	0.22 ± 0.04	$1.73, > 0.99$
nnU-Net FL	PLAn 2 FL/T1w	0.13 ± 0.04	0.20 ± 0.05	$2.73, 0.179$	0.13 ± 0.04	0.21 ± 0.05	$1.42, > 0.99$
nnU-Net FL/T1w	PLAn 1 FL	0.10 ± 0.04	0.20 ± 0.05	$3.36, 0.022$	0.10 ± 0.04	0.20 ± 0.05	$2.29, 0.788$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.10 ± 0.04	0.18 ± 0.05	$2.44, 0.415$	0.10 ± 0.04	0.18 ± 0.05	$1.06, > 0.99$
nnU-Net FL/T1w	PLAn 2 FL	0.10 ± 0.04	0.21 ± 0.04	$3.47, 0.015$	0.10 ± 0.04	0.22 ± 0.04	$2.25, 0.886$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.10 ± 0.04	0.20 ± 0.05	$3.84, 0.003$	0.10 ± 0.04	0.21 ± 0.05	$1.93, > 0.99$
PLAn 1 FL	PLAn 1 FL/T1w	0.20 ± 0.05	0.18 ± 0.05	$0.92, > 0.99$	0.20 ± 0.05	0.18 ± 0.05	$1.24, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.20 ± 0.05	0.21 ± 0.04	$0.11, > 0.99$	0.20 ± 0.05	0.22 ± 0.04	$0.04, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.20 ± 0.05	0.20 ± 0.05	$0.48, > 0.99$	0.20 ± 0.05	0.21 ± 0.05	$0.36, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.18 ± 0.05	0.21 ± 0.04	$1.03, > 0.99$	0.18 ± 0.05	0.22 ± 0.04	$1.19, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.18 ± 0.05	0.20 ± 0.05	$1.40, > 0.99$	0.18 ± 0.05	0.21 ± 0.05	$0.88, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.21 ± 0.04	0.20 ± 0.05	$0.37, > 0.99$	0.22 ± 0.04	0.21 ± 0.05	$0.31, > 0.99$

Abbreviations: FPR = false positive ratio; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.12: FNR cluster pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference in ULF Blinded Subset. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.38 ± 0.05	0.31 ± 0.04	$1.55, > 0.99$	0.36 ± 0.05	0.29 ± 0.04	$0.16, > 0.99$
MiMoSA	nnU-Net FL	0.38 ± 0.05	0.59 ± 0.05	$7.29, < 0.001$	0.36 ± 0.05	0.60 ± 0.05	$5.55, < 0.001$
MiMoSA	nnU-Net FL/T1w	0.38 ± 0.05	0.60 ± 0.05	$7.83, < 0.001$	0.36 ± 0.05	0.60 ± 0.05	$5.71, < 0.001$
MiMoSA	PLAn 1 FL	0.38 ± 0.05	0.50 ± 0.04	$3.63, 0.008$	0.36 ± 0.05	0.50 ± 0.04	$3.15, 0.059$
MiMoSA	PLAn 1 FL/T1w	0.38 ± 0.05	0.47 ± 0.05	$3.78, 0.004$	0.36 ± 0.05	0.47 ± 0.05	$2.61, 0.329$
MiMoSA	PLAn 2 FL	0.38 ± 0.05	0.49 ± 0.05	$3.63, 0.008$	0.36 ± 0.05	0.49 ± 0.05	$2.79, 0.192$
MiMoSA	PLAn 2 FL/T1w	0.38 ± 0.05	0.48 ± 0.05	$3.92, 0.003$	0.36 ± 0.05	0.48 ± 0.05	$2.56, 0.375$
SynthSeg	nnU-Net FL	0.31 ± 0.04	0.59 ± 0.05	$5.75, < 0.001$	0.29 ± 0.04	0.60 ± 0.05	$5.71, < 0.001$
SynthSeg	nnU-Net FL/T1w	0.31 ± 0.04	0.60 ± 0.05	$6.28, < 0.001$	0.29 ± 0.04	0.60 ± 0.05	$5.87, < 0.001$
SynthSeg	PLAn 1 FL	0.31 ± 0.04	0.50 ± 0.04	$2.08, > 0.99$	0.29 ± 0.04	0.50 ± 0.04	$3.30, 0.034$
SynthSeg	PLAn 1 FL/T1w	0.31 ± 0.04	0.47 ± 0.05	$2.23, 0.727$	0.29 ± 0.04	0.47 ± 0.05	$2.76, 0.205$
SynthSeg	PLAn 2 FL	0.31 ± 0.04	0.49 ± 0.05	$2.08, > 0.99$	0.29 ± 0.04	0.49 ± 0.05	$2.94, 0.117$
SynthSeg	PLAn 2 FL/T1w	0.31 ± 0.04	0.48 ± 0.05	$2.37, 0.495$	0.29 ± 0.04	0.48 ± 0.05	$2.72, 0.235$
nnU-Net FL	nnU-Net FL/T1w	0.59 ± 0.05	0.60 ± 0.05	$0.53, > 0.99$	0.60 ± 0.05	0.60 ± 0.05	$0.16, > 0.99$
nnU-Net FL	PLAn 1 FL	0.59 ± 0.05	0.50 ± 0.04	$3.66, 0.007$	0.60 ± 0.05	0.50 ± 0.04	$2.40, 0.583$
nnU-Net FL	PLAn 1 FL/T1w	0.59 ± 0.05	0.47 ± 0.05	$3.52, 0.012$	0.60 ± 0.05	0.47 ± 0.05	$2.94, 0.117$
nnU-Net FL	PLAn 2 FL	0.59 ± 0.05	0.49 ± 0.05	$3.66, 0.007$	0.60 ± 0.05	0.49 ± 0.05	$2.76, 0.205$
nnU-Net FL	PLAn 2 FL/T1w	0.59 ± 0.05	0.48 ± 0.05	$3.37, 0.021$	0.60 ± 0.05	0.48 ± 0.05	$2.99, 0.101$
nnU-Net FL/T1w	PLAn 1 FL	0.60 ± 0.05	0.50 ± 0.04	$4.20, < 0.001$	0.60 ± 0.05	0.50 ± 0.04	$2.56, 0.375$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.60 ± 0.05	0.47 ± 0.05	$4.05, 0.001$	0.60 ± 0.05	0.47 ± 0.05	$3.10, 0.069$
nnU-Net FL/T1w	PLAn 2 FL	0.60 ± 0.05	0.49 ± 0.05	$4.20, < 0.001$	0.60 ± 0.05	0.49 ± 0.05	$2.92, 0.125$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.60 ± 0.05	0.48 ± 0.05	$3.90, 0.003$	0.60 ± 0.05	0.48 ± 0.05	$3.15, 0.059$
PLAn 1 FL	PLAn 1 FL/T1w	0.50 ± 0.04	0.47 ± 0.05	$0.15, > 0.99$	0.50 ± 0.04	0.47 ± 0.05	$0.54, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.50 ± 0.04	0.49 ± 0.05	$0.00, > 0.99$	0.50 ± 0.04	0.49 ± 0.05	$0.36, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.50 ± 0.04	0.48 ± 0.05	$0.29, > 0.99$	0.50 ± 0.04	0.48 ± 0.05	$0.58, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.47 ± 0.05	0.49 ± 0.05	$0.15, > 0.99$	0.47 ± 0.05	0.49 ± 0.05	$0.18, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.47 ± 0.05	0.48 ± 0.05	$0.15, > 0.99$	0.47 ± 0.05	0.48 ± 0.05	$0.04, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.49 ± 0.05	0.48 ± 0.05	$0.29, > 0.99$	0.49 ± 0.05	0.48 ± 0.05	$0.22, > 0.99$

Abbreviations: FNR = false negative ratio; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.13: DER pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference in ULF Blinded Subset. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	1.34 ± 0.10	0.52 ± 0.07	$7.55, < 0.001$	1.32 ± 0.10	0.31 ± 0.06	$6.41, < 0.001$
MiMoSA	nnU-Net FL	1.34 ± 0.10	0.76 ± 0.13	$6.57, < 0.001$	1.32 ± 0.10	0.75 ± 0.13	$3.15, 0.059$
MiMoSA	nnU-Net FL/T1w	1.34 ± 0.10	0.72 ± 0.13	$6.84, < 0.001$	1.32 ± 0.10	0.71 ± 0.13	$3.39, 0.025$
MiMoSA	PLAn 1 FL	1.34 ± 0.10	0.57 ± 0.10	$10.15, < 0.001$	1.32 ± 0.10	0.56 ± 0.10	$5.55, < 0.001$
MiMoSA	PLAn 1 FL/T1w	1.34 ± 0.10	0.54 ± 0.11	$11.09, < 0.001$	1.32 ± 0.10	0.54 ± 0.11	$7.10, < 0.001$
MiMoSA	PLAn 2 FL	1.34 ± 0.10	0.58 ± 0.11	$10.83, < 0.001$	1.32 ± 0.10	0.58 ± 0.11	$6.18, < 0.001$
MiMoSA	PLAn 2 FL/T1w	1.34 ± 0.10	0.54 ± 0.11	$10.75, < 0.001$	1.32 ± 0.10	0.53 ± 0.11	$7.03, < 0.001$
SynthSeg	nnU-Net FL	0.52 ± 0.07	0.76 ± 0.13	$0.98, > 0.99$	0.31 ± 0.06	0.75 ± 0.13	$3.26, 0.040$
SynthSeg	nnU-Net FL/T1w	0.52 ± 0.07	0.72 ± 0.13	$0.71, > 0.99$	0.31 ± 0.06	0.71 ± 0.13	$3.01, 0.094$
SynthSeg	PLAn 1 FL	0.52 ± 0.07	0.57 ± 0.10	$2.60, 0.263$	0.31 ± 0.06	0.56 ± 0.10	$0.85, > 0.99$
SynthSeg	PLAn 1 FL/T1w	0.52 ± 0.07	0.54 ± 0.11	$3.53, 0.011$	0.31 ± 0.06	0.54 ± 0.11	$0.70, > 0.99$
SynthSeg	PLAn 2 FL	0.52 ± 0.07	0.58 ± 0.11	$3.28, 0.029$	0.31 ± 0.06	0.58 ± 0.11	$0.22, > 0.99$
SynthSeg	PLAn 2 FL/T1w	0.52 ± 0.07	0.54 ± 0.11	$3.19, 0.039$	0.31 ± 0.06	0.53 ± 0.11	$0.63, > 0.99$
nnU-Net FL	nnU-Net FL/T1w	0.76 ± 0.13	0.72 ± 0.13	$0.27, > 0.99$	0.75 ± 0.13	0.71 ± 0.13	$0.25, > 0.99$
nnU-Net FL	PLAn 1 FL	0.76 ± 0.13	0.57 ± 0.10	$3.58, 0.009$	0.75 ± 0.13	0.56 ± 0.10	$2.40, 0.583$
nnU-Net FL	PLAn 1 FL/T1w	0.76 ± 0.13	0.54 ± 0.11	$4.52, < 0.001$	0.75 ± 0.13	0.54 ± 0.11	$3.96, 0.003$
nnU-Net FL	PLAn 2 FL	0.76 ± 0.13	0.58 ± 0.11	$4.26, < 0.001$	0.75 ± 0.13	0.58 ± 0.11	$3.03, 0.087$
nnU-Net FL	PLAn 2 FL/T1w	0.76 ± 0.13	0.54 ± 0.11	$4.18, < 0.001$	0.75 ± 0.13	0.53 ± 0.11	$3.89, 0.004$
nnU-Net FL/T1w	PLAn 1 FL	0.72 ± 0.13	0.57 ± 0.10	$3.31, 0.026$	0.71 ± 0.13	0.56 ± 0.10	$2.16, > 0.99$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.72 ± 0.13	0.54 ± 0.11	$4.24, < 0.001$	0.71 ± 0.13	0.54 ± 0.11	$3.71, 0.007$
nnU-Net FL/T1w	PLAn 2 FL	0.72 ± 0.13	0.58 ± 0.11	$3.99, 0.002$	0.71 ± 0.13	0.58 ± 0.11	$2.79, 0.192$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.72 ± 0.13	0.54 ± 0.11	$3.90, 0.003$	0.71 ± 0.13	0.53 ± 0.11	$3.64, 0.010$
PLAn 1 FL	PLAn 1 FL/T1w	0.57 ± 0.10	0.54 ± 0.11	$0.94, > 0.99$	0.56 ± 0.10	0.54 ± 0.11	$1.55, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.57 ± 0.10	0.58 ± 0.11	$0.68, > 0.99$	0.56 ± 0.10	0.58 ± 0.11	$0.63, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.57 ± 0.10	0.54 ± 0.11	$0.60, > 0.99$	0.56 ± 0.10	0.53 ± 0.11	$1.48, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.54 ± 0.11	0.58 ± 0.11	$0.26, > 0.99$	0.54 ± 0.11	0.58 ± 0.11	$0.92, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.54 ± 0.11	0.54 ± 0.11	$0.34, > 0.99$	0.54 ± 0.11	0.53 ± 0.11	$0.07, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.58 ± 0.11	0.54 ± 0.11	$0.08, > 0.99$	0.58 ± 0.11	0.53 ± 0.11	$0.85, > 0.99$

Abbreviations: DER = detection error rate; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

Table D.14: OER pairwise post hoc tests (Dunn's) between automatic ULF WML segmentation methods, with ULF Aided masks as reference in ULF Blinded Subset. Analyses were performed both in the original ULF native space/resolution and in resampled 1×1×1 mm^3 space.

		Original Space			**1×1×1 mm^3 Space**		
Method 1	**Method 2**	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*	*Mean ± SE (M1)*	*Mean ± SE (M2)*	*Z, adj. p*
MiMoSA	SynthSeg	0.25 ± 0.04	0.89 ± 0.04	$13.52, < 0.001$	0.26 ± 0.04	1.13 ± 0.06	$9.33, < 0.001$
MiMoSA	nnU-Net FL	0.25 ± 0.04	0.50 ± 0.06	$7.78, < 0.001$	0.26 ± 0.04	0.50 ± 0.06	$3.48, 0.018$
MiMoSA	nnU-Net FL/T1w	0.25 ± 0.04	0.53 ± 0.06	$6.83, < 0.001$	0.26 ± 0.04	0.54 ± 0.06	$4.04, 0.002$
MiMoSA	PLAn 1 FL	0.25 ± 0.04	0.57 ± 0.05	$5.34, < 0.001$	0.26 ± 0.04	0.58 ± 0.05	$3.62, 0.011$
MiMoSA	PLAn 1 FL/T1w	0.25 ± 0.04	0.55 ± 0.05	$6.10, < 0.001$	0.26 ± 0.04	0.55 ± 0.05	$3.87, 0.004$
MiMoSA	PLAn 2 FL	0.25 ± 0.04	0.53 ± 0.04	$5.70, < 0.001$	0.26 ± 0.04	0.53 ± 0.04	$3.64, 0.010$
MiMoSA	PLAn 2 FL/T1w	0.25 ± 0.04	0.57 ± 0.05	$6.63, < 0.001$	0.26 ± 0.04	0.57 ± 0.05	$4.02, 0.002$
SynthSeg	nnU-Net FL	0.89 ± 0.04	0.50 ± 0.06	$5.75, < 0.001$	1.13 ± 0.06	0.50 ± 0.06	$5.84, < 0.001$
SynthSeg	nnU-Net FL/T1w	0.89 ± 0.04	0.53 ± 0.06	$6.70, < 0.001$	1.13 ± 0.06	0.54 ± 0.06	$5.28, < 0.001$
SynthSeg	PLAn 1 FL	0.89 ± 0.04	0.57 ± 0.05	$8.18, < 0.001$	1.13 ± 0.06	0.58 ± 0.05	$5.71, < 0.001$
SynthSeg	PLAn 1 FL/T1w	0.89 ± 0.04	0.55 ± 0.05	$7.42, < 0.001$	1.13 ± 0.06	0.55 ± 0.05	$5.46, < 0.001$
SynthSeg	PLAn 2 FL	0.89 ± 0.04	0.53 ± 0.04	$7.83, < 0.001$	1.13 ± 0.06	0.53 ± 0.04	$5.69, < 0.001$
SynthSeg	PLAn 2 FL/T1w	0.89 ± 0.04	0.57 ± 0.05	$6.89, < 0.001$	1.13 ± 0.06	0.57 ± 0.05	$5.30, < 0.001$
nnU-Net FL	nnU-Net FL/T1w	0.50 ± 0.06	0.53 ± 0.06	$0.95, > 0.99$	0.50 ± 0.06	0.54 ± 0.06	$0.56, > 0.99$
nnU-Net FL	PLAn 1 FL	0.50 ± 0.06	0.57 ± 0.05	$2.44, 0.415$	0.50 ± 0.06	0.58 ± 0.05	$0.13, > 0.99$
nnU-Net FL	PLAn 1 FL/T1w	0.50 ± 0.06	0.55 ± 0.05	$1.68, > 0.99$	0.50 ± 0.06	0.55 ± 0.05	$0.38, > 0.99$
nnU-Net FL	PLAn 2 FL	0.50 ± 0.06	0.53 ± 0.04	$2.08, > 0.99$	0.50 ± 0.06	0.53 ± 0.04	$0.16, > 0.99$
nnU-Net FL	PLAn 2 FL/T1w	0.50 ± 0.06	0.57 ± 0.05	$1.15, > 0.99$	0.50 ± 0.06	0.57 ± 0.05	$0.54, > 0.99$
nnU-Net FL/T1w	PLAn 1 FL	0.53 ± 0.06	0.57 ± 0.05	$1.49, > 0.99$	0.54 ± 0.06	0.58 ± 0.05	$0.43, > 0.99$
nnU-Net FL/T1w	PLAn 1 FL/T1w	0.53 ± 0.06	0.55 ± 0.05	$0.73, > 0.99$	0.54 ± 0.06	0.55 ± 0.05	$0.18, > 0.99$
nnU-Net FL/T1w	PLAn 2 FL	0.53 ± 0.06	0.53 ± 0.04	$1.13, > 0.99$	0.54 ± 0.06	0.53 ± 0.04	$0.40, > 0.99$
nnU-Net FL/T1w	PLAn 2 FL/T1w	0.53 ± 0.06	0.57 ± 0.05	$0.19, > 0.99$	0.54 ± 0.06	0.57 ± 0.05	$0.02, > 0.99$
PLAn 1 FL	PLAn 1 FL/T1w	0.57 ± 0.05	0.55 ± 0.05	$0.76, > 0.99$	0.58 ± 0.05	0.55 ± 0.05	$0.25, > 0.99$
PLAn 1 FL	PLAn 2 FL	0.57 ± 0.05	0.53 ± 0.04	$0.35, > 0.99$	0.58 ± 0.05	0.53 ± 0.04	$0.02, > 0.99$
PLAn 1 FL	PLAn 2 FL/T1w	0.57 ± 0.05	0.57 ± 0.05	$1.29, > 0.99$	0.58 ± 0.05	0.57 ± 0.05	$0.40, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL	0.55 ± 0.05	0.53 ± 0.04	$0.40, > 0.99$	0.55 ± 0.05	0.53 ± 0.04	$0.22, > 0.99$
PLAn 1 FL/T1w	PLAn 2 FL/T1w	0.55 ± 0.05	0.57 ± 0.05	$0.53, > 0.99$	0.55 ± 0.05	0.57 ± 0.05	$0.16, > 0.99$
PLAn 2 FL	PLAn 2 FL/T1w	0.53 ± 0.04	0.57 ± 0.05	$0.94, > 0.99$	0.53 ± 0.04	0.57 ± 0.05	$0.38, > 0.99$

Abbreviations: OER = outline error rate; ULF = ultra-low field (64 mT); WML = white matter lesions; SE = standard error; adj. p = adjusted p-value; FL = fluid-attenuated inversion recovery; T1w = T1-weighted;

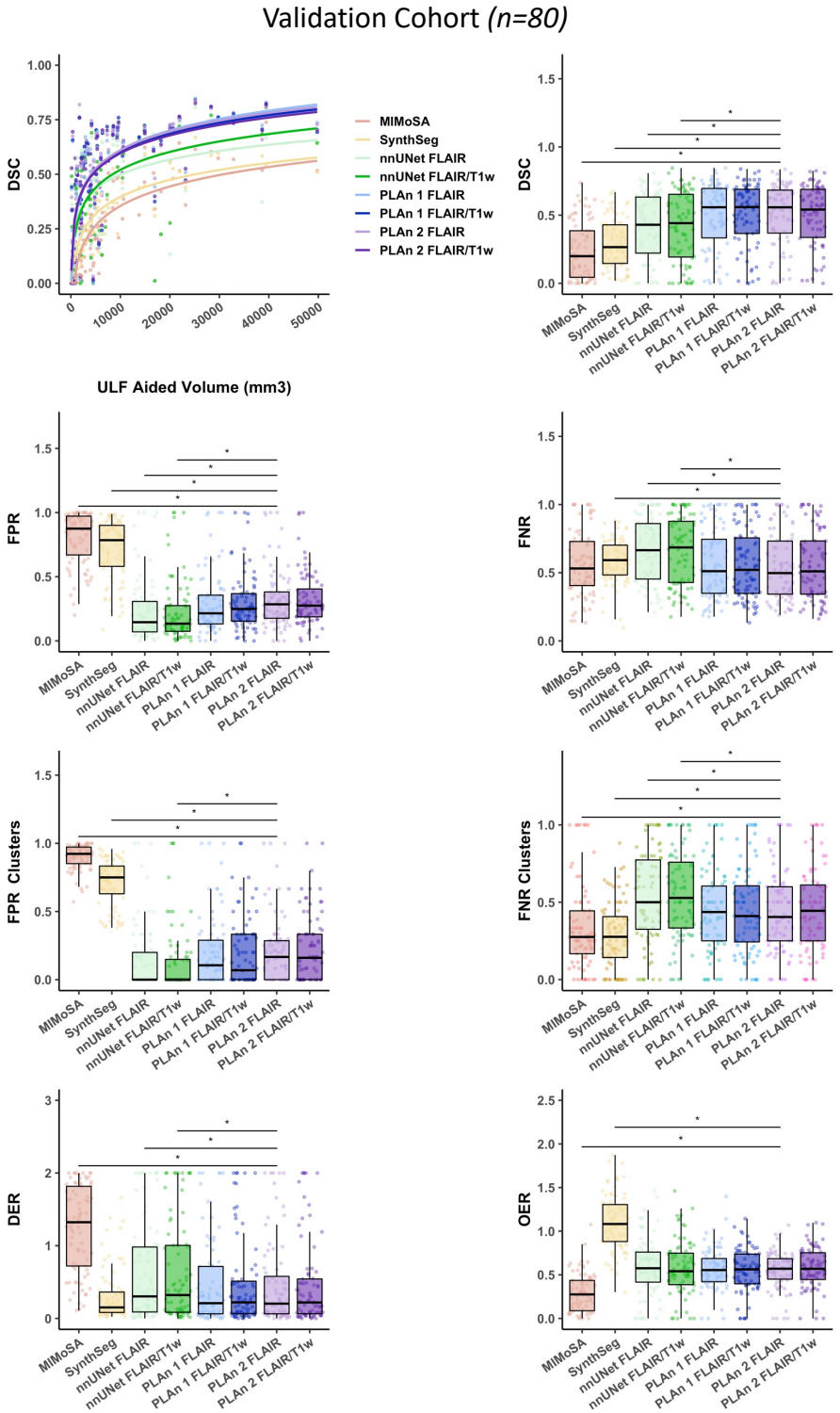


Figure D.1: Evaluation of automatic WML segmentation performance in $1 \times 1 \times 1$ mm^3 resolution. Overlap metrics of automatic segmentation algorithms using ULF Aided Manual masks as a reference. Scatter plots of DSC vs. ULF Aided volume are fitted with logarithmic non-linear regression for each method. A) Overlap metrics in the entire validation cohort ($n = 80$). Significant differences are only shown between PLAn 2 FLAIR and other methods.

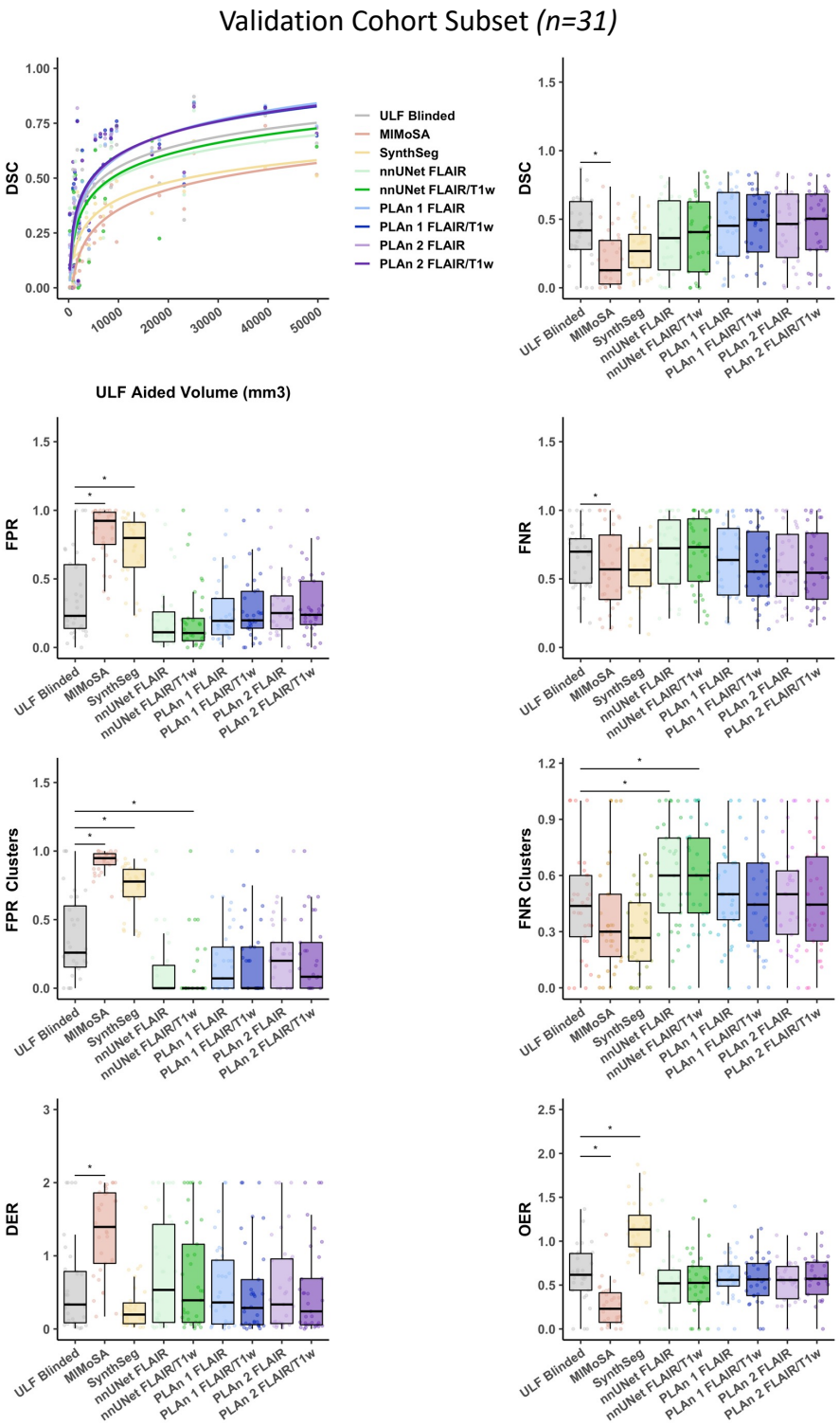


Figure D.2: Evaluation of automatic WML segmentation performance in $1 \times 1 \times 1$ mm^3 resolution. Overlap metrics of automatic segmentation algorithms using ULF Aided Manual masks as a reference. Scatter plots of DSC vs. ULF Aided volume are fitted with logarithmic non-linear regression for each method. Overlap metrics in the subset of the validation cohort with ULF Blinded masks ($n = 31$). Significant differences are only shown between ULF Blinded and other methods.

D.1.2 Performance of Automatic Segmentation Algorithm across ULF Acquisition Parameters

In this appendix I present the results of the exploratory analysis of ULF automatic segmentation performance across heterogeneous acquisition parameters (Summarised in the main text Section 4.1.3.5).

The validation dataset included scans with varying reconstruction methods, software versions, and FLAIR in-plane voxel sizes (Table 4.2). Due to the lack of a controlled dataset, direct comparisons of different acquisition settings were not possible. Instead, I evaluated whether general DSC trends were consistent across participant subsets with specific ULF acquisition parameters.

Most of the validation cohort had MRI scans reconstructed using a Cartesian method ($n = 66$), while only 18 cases used deep learning (DL) reconstruction. Given that total WML volume influences DSC, with higher WML volumes generally associated with improved segmentation performance, I first compared WML volumes between the Cartesian and DL cohorts (Figure D.3A, left). No significant differences were observed in 3T-measured lesion volumes between the two groups (Cartesian: 5651 ± 908 mm^3, DL: 8718 ± 3075 mm^3, $U = 558$, $p = 0.7$, two-tailed), nor in ULF-measured lesion volumes (Cartesian: 7776 ± 1199 mm^3, DL: 9050 ± 2872 mm^3, $U = 575$, $p = 0.8$, two-tailed). This suggests that any observed differences in segmentation performance are unlikely to be driven by differences in total lesion burden.

Across both the DL and cartesian cohorts, all PLAn models demonstrated significantly higher DSC scores than MIMoSA, indicating superior segmentation performance. However, nnU-Net models showed reduced segmentation performance in the DL cohort, while other methods maintained consistent performance between DL and cartesian scans (Figure D.3A). This suggests that nnU-Net may be more sensitive to reconstruction method variability than PLAn or SynthSeg.

A Kruskal-Wallis test showed no significant differences in either 3T lesion volumes ($H = 1.2$, $p = 0.09$) or ULF Aided lesion volumes ($H = 6.0$, $p = 0.12$) across ULF software versions (Figure D.3B, left). Similarly, no significant differences were found in 3T volumes ($H = 2.5$, $p = 0.57$) or ULF volumes ($H = 1.7$, $p = 0.67$) across ULF in-plane voxel sizes (Figure D.3C, left).

Segmentation performance among PLAn models remained relatively stable across ULF software versions and voxel sizes, consistently achieving the highest mean DSC scores in all sub-cohorts, with no significant differences observed

between them. In contrast, nnU-Net models exhibited the greatest variability in performance, achieving higher DSC scores in newer software versions and at a 1.7 mm resolution. Notably, these specific scan settings were the most common in the training dataset, which may explain the improved performance of nnU-Net under these conditions. SynthSeg performed consistently across all sub-cohorts, with DSC scores between MIMoSA and PLAn models, demonstrating greater robustness to scan parameter variations than nnU-Net but lower overall segmentation accuracy than PLAn (Figure D.3).

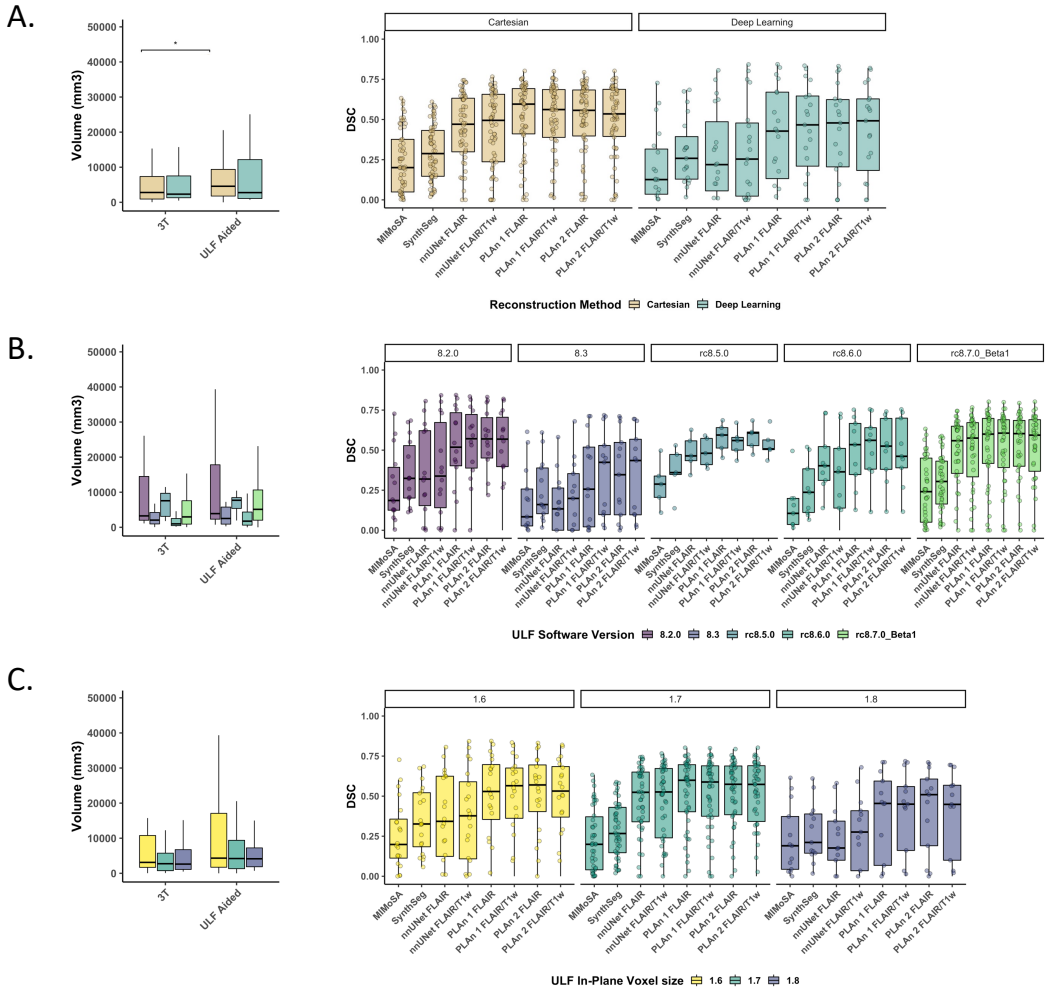


Figure D.3: 3T and ULF aided Volume (left) and automatic segmentation DSC (right) in Validation dataset with the dataset split based on ULF A) reconstruction method, B) ULF software version, and C) ULF-In plane voxel size. Significant differences are only indicated for the volume plots (left).

D.2 Supplementary data for Chapter 4.2

D.2.1 Normalisation during ULF SUBLONG Pipeline Development

Given the variability in ULF acquisition parameters between timepoints, *WhiteStripe* normalisation was selected for intensity normalisation. This approach was preferred because it more effectively produced consistent intensity profiles across longitudinal scans compared to other methods. As shown in figure D.4, *WhiteStripe* normalisation yielded better alignment of intensity distributions between baseline and follow-up timepoints than a simple 98th percentile normalisation.

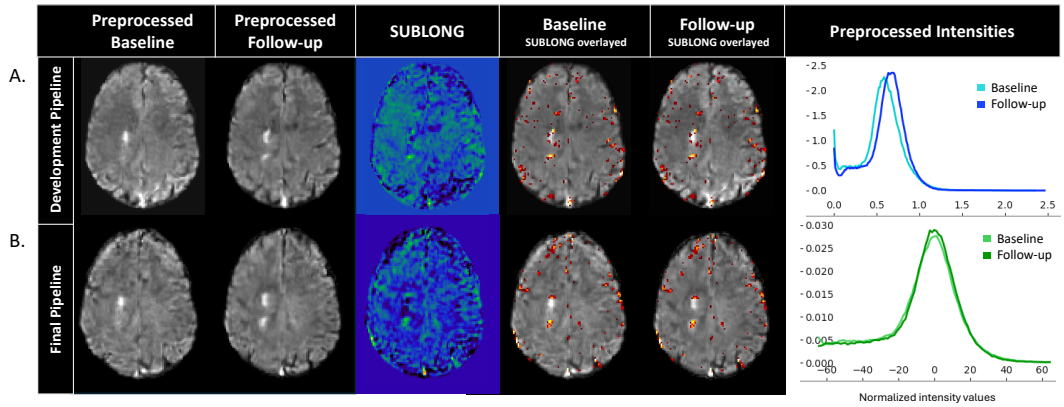


Figure D.4: Comparison of (from left to right) longitudinal preprocessed ULF FLAIR scans at baseline and follow-up, ULF SUBLONG outputs (shown alone and overlaid on the preprocessed ULF FLAIRs), and intensity distributions of preprocessed ULF FLAIR at baseline and follow-up. A) Early version of the ULF SUBLONG pipeline developed using 98th percentile normalisation. B) Final version of the ULF SUBLONG pipeline using *WhiteStripe* normalisation.

D.2.2 Impact of ULF Acquisition Parameters on ULF SUBLONG Analysis

In this appendix I present the results of the exploratory analysis of ULF WML evolution detection across heterogeneous acquisition parameters (Summarised in the main text Section 4.2.3.3).

Variability in ULF scan parameters may impact ULF SUBLONG quality and radiological detection of CLs and NLs, and may contribute to discrepancies in lesion volume measurements between ULF and 3T. To explore these potential effects, three ULF acquisition parameters were assessed:

- Reconstruction method: deep learning (DL) or Cartesian (cart)
- Hyperfine Software version

- In-plane voxel size

Table D.15 provides an overview of ULF reconstruction method, in-plane voxel size and Hyperfine software for each participant's baseline and follow-up scan and related metrics that may have influenced the results.

Table D.15: Table Heterogeneity in ULF Acquisition Parameters Between Baseline and Follow-Up (Reconstruction Method, In-Plane Voxel Size, Software Version). Each row represents one participant.

	ULF Recon. Method			**ULF In-Plane Voxel Size**				**Hyperfine Software Version**			
≥ 1 NL or CL	**TP 1**	**TP 2**	**Recon Δ**	**TP 1**	**TP 2**	**Voxel Size Δ**	**Avg. Voxel Size**	**TP 1**	**TP 2**	**Soft. Ver. Δ**	**Soft. Ver. 8.2**
No	DL	cart	change	1.6	1.7	0.1	1.65	rc8.2.0	rc8.6.0	0.4	Yes
No	DL	cart	change	1.6	1.8	0.2	1.7	rc8.2.0	rc8.5.0	0.3	Yes
No	DL	DL	No change	1.6	1.8	0.2	1.7	rc8.2.0	rc8.3.0	0.1	Yes
No	DL	cart	change	1.6	1.7	0.1	1.65	8.2.0	rc8.7.0	0.501	Yes
No	cart	cart	change	1.8	1.8	0	1.8	rc8.3.0	8.3.0	0	No
No	cart	cart	No change	1.8	1.6	-0.2	1.7	rc8.3.2	rc8.5.0	0.18	No
No	cart	cart	No change	1.6	1.7	0.1	1.65	rc8.3.2	rc8.6.0	0.28	No
No	cart	cart	No change	1.8	1.7	-0.1	1.75	rc8.5.0	rc8.7.0	0.201	No
No	cart	cart	No change	1.7	1.7	0	1.7	rc8.6.0	rc8.7.0	0.101	No
Yes	cart	cart	No change	1.6	1.7	0.1	1.65	rc8.1.1	rc8.7.0	0.591	No
Yes	DL	cart	change	1.6	1.7	0.1	1.65	rc8.2.0	rc8.7.0	0.501	Yes
Yes	DL	cart	change	1.6	1.8	0.2	1.7	rc8.2.0	rc8.3.2	0.12	Yes
Yes	DL	cart	change	1.6	1.8	0.2	1.7	rc8.2.0	8.3.2	0.12	Yes
Yes	DL	cart	change	1.6	1.8	0.2	1.7	rc8.2.0	8.3.2	0.12	Yes
Yes	DL	cart	change	1.6	1.7	0.1	1.65	8.2.0	rc8.7.0	0.501	Yes
Yes	DL	cart	change	1.8	1.7	-0.1	1.75	rc8.3.0	rc8.7.0	0.401	No
Yes	cart	cart	No change	1.8	1.7	-0.1	1.75	8.3.0	rc8.7.0	0.401	No
Yes	cart	cart	No change	1.8	1.7	-0.1	1.75	8.3.0	rc8.6.0	0.3	No
Yes	cart	cart	No change	1.8	1.7	-0.1	1.75	8.3.0	rc8.7.0	0.401	No
Yes	cart	cart	No change	1.8	1.7	-0.1	1.75	rc8.3.2	rc8.6.0	0.28	No
Yes	cart	cart	No change	1.8	1.7	-0.1	1.75	rc8.3.2	rc8.7.0	0.381	No
Yes	cart	cart	No change	1.8	1.6	-0.2	1.7	8.3.2	rc8.5.0	0.18	No
Yes	cart	cart	No change	1.8	1.7	-0.1	1.75	8.3.2	rc8.6.0	0.28	No

Abbreviations: NL = New Lesion; CL = Changing Lesion; TP 1 = timepoint 1 (baseline); TP 2 = timepoint 1 (follow-up); Recon = reconstruction method; Δ = change from baseline to follow-up; Avg. = average of baseline and follow-up values; Soft. Ver. = software version;

For each participant with at least one CL or NL, the percent of changing (Figure D.5) or new (Figure D.6) lesions detected at ULF was plotted against several acquisition-related metrics:

- Change in Reconstruction Method: Whether the reconstruction method differed between baseline and follow-up MRI scans (e.g., *DL-Cart* indicates DL reconstruction at baseline and Cartesian reconstruction at follow-up).
- Change in In-Plane Voxel Size: The difference in in-plane voxel size between follow-up and baseline scans.
- Average In-Plane Voxel Size: The average in-plane voxel size calculated from baseline and follow-up scans.
- Amount of Software Version Change: Calculated as the numerical difference between baseline and follow-up software versions (e.g., rc8.6.0 and 8.2.0 corresponds to a 0.4 change). Smaller values indicate scans acquired with software versions released closer together in time, though not necessarily more similar in function.
- Contains Version 8.2: Indicates if the participant had at least one scan (baseline or follow-up) acquired with Hyperfine software version 8.2, which is anecdotally linked to lower image quality.

There were no clear patterns of acquisition parameters on percent detection of CL on ULF. However, participants with undetected CLs often had either a change in reconstruction method between timepoints or scans acquired with software version 8.3.2, which is known to degrade FLAIR image quality (Appendix D.5).

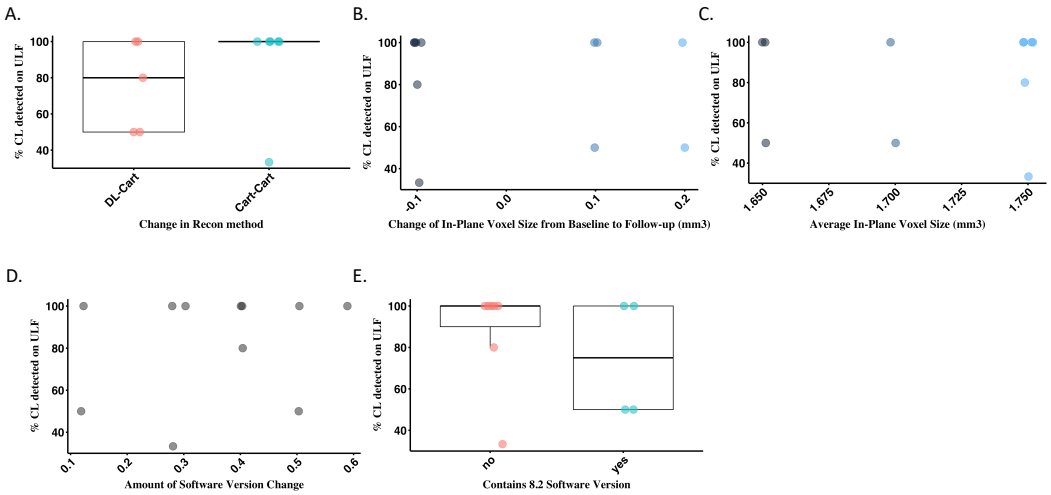


Figure D.5: Participant-level percent detection of changing lesions (CLs) at ULF compared to 3T by ULF acquisition related metrics. Percent of CLs detected by: A) change in reconstruction method, B) voxel size difference, C) average voxel size, D) software version change, and E) presence of version 8.2.

Similarly, there were no clear patterns of acquisition parameters on percent detection of NL on ULF (Appendix D.6).

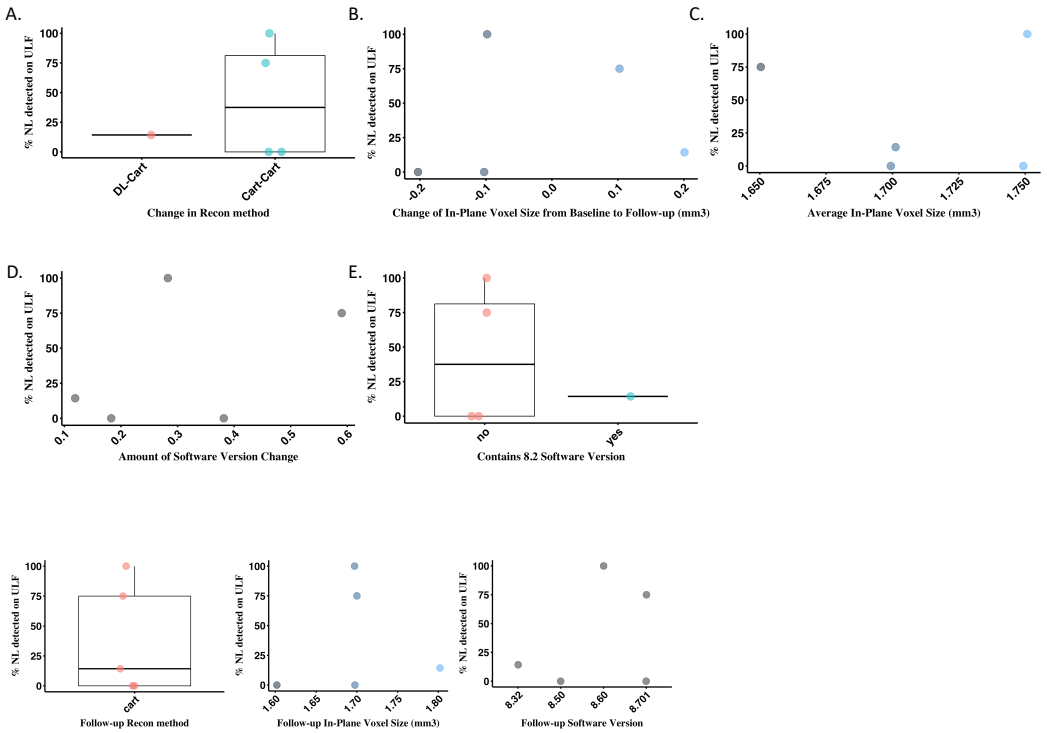


Figure D.6: Participant-level percent detection of new lesions (NLs) at ULF compared to 3T by ULF acquisition related metrics. Percent of NLs detected by: A) change in reconstruction method, B) voxel size difference, C) average voxel size, D) software version change, and E) presence of version 8.2.

CLs or NLs were pooled across all participants with at least one CL or NL, respectively. ULF acquisition-related metrics were compared for CLs or NLs radiologically detected vs missed at ULF. The distribution and number of CLs or NLs detected on ULF compared to those not detected on ULF didn't show clear associations with ULF Acquisition (Figure D.7 and D.8, respectively). Importantly, in figure D.8A and figure D.8E, only one participant had DL-Cart reconstruction change and a Hyperfine 8.2 Software Version. This participant had 5 NLs on 3T one of which was detected at ULF.

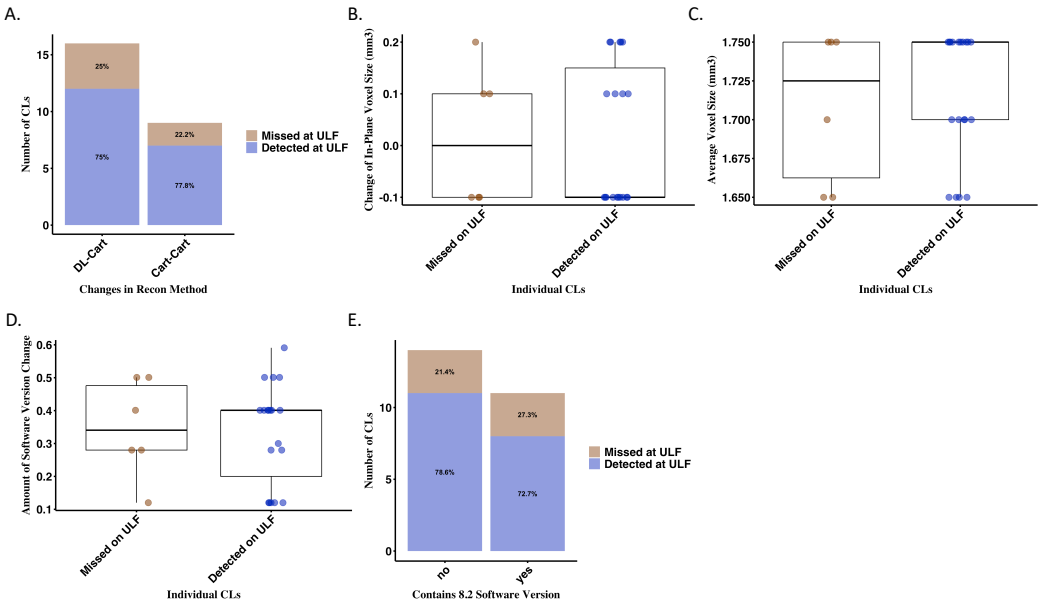


Figure D.7: Impact of ULF acquisition parameters on the individual changing lesions (CLs) radiologically detected vs missed at ULF ($n = 23$). A) Number of CLs detected vs missed on ULF for participants with or without a change in reconstruction method. B) Voxel size difference- C) Average voxel size- D) and magnitude of software version change for CLs detected vs missed at ULF. E) Number of CLs detected vs missed on ULF for participants with or without at least one MRI image scanned using Hyperfine Software Version 8.2.

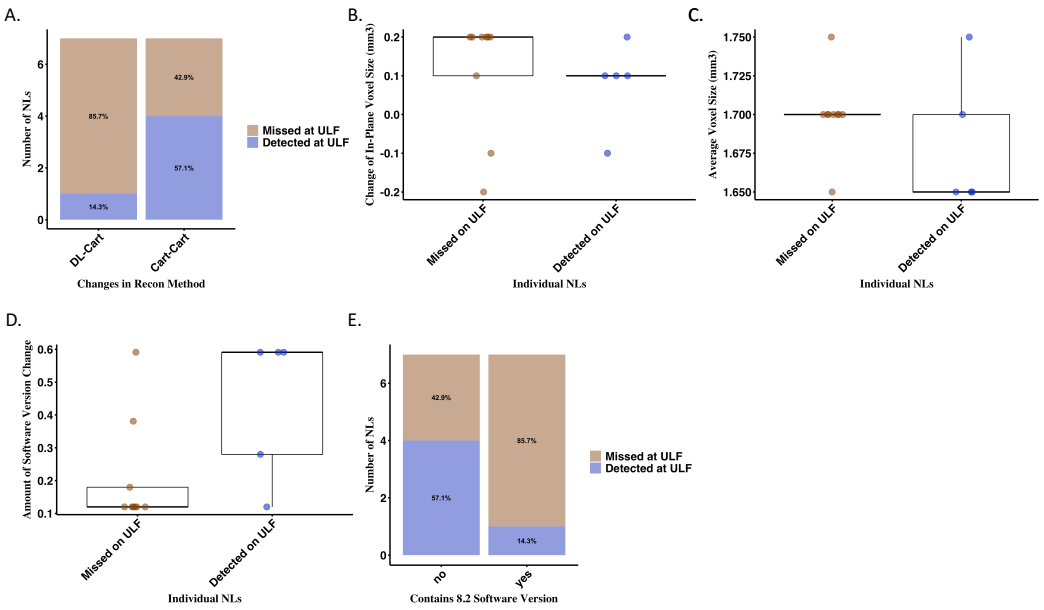


Figure D.8: Impact of ULF acquisition parameters on the individual new lesions (NLs) radiologically detected vs missed at ULF ($n = 23$). A) Number of NLs detected vs missed on ULF for participants with or without a change in reconstruction method. B) Voxel size difference- C) Average voxel size- D) and magnitude of software version change for NLs detected vs missed at ULF. E) Number of NLs detected vs missed on ULF for participants with or without at least one MRI image scanned using Hyperfine Software Version 8.2.

Finally, ULF acquisition parameters were assessed in relation to discrepan-

cies in lesion volume measurements between ULF and 3T. CLs and NLs were pooled across all participants with at least one CL or NL, respectively. For CLs, the change in lesion volume from baseline to follow-up was calculated at both 3T and ULF. ULF acquisition-related metrics were then compared to the difference in 3T and ULF volume change. The CL with the greatest discrepancy between 3T and ULF volume change had a change in reconstruction method between baseline and follow-up (Figure D.9) and included the 8.2 software version in at least one of the acquisitions. After removing this case (given its large ULF–3T difference), CLs with DL at baseline and Cartesian at follow-up reconstruction tended to show greater discrepancies in 3T–ULF volume change than cases with Cartesian reconstruction at both baseline and follow-up (Figure D.10A). Similarly, CLs with at least one timepoint acquired using software version 8.2 tended to show greater discrepancies in 3T–ULF volume change than cases without this version at either timepoint (Figure D.10E).

For NLs, lesion volume at follow-up was calculated at both 3T and ULF. ULF acquisition parameters at follow-up were then compared to the difference in 3T and ULF NL volume. No clear trends were observed (Figure D.11).

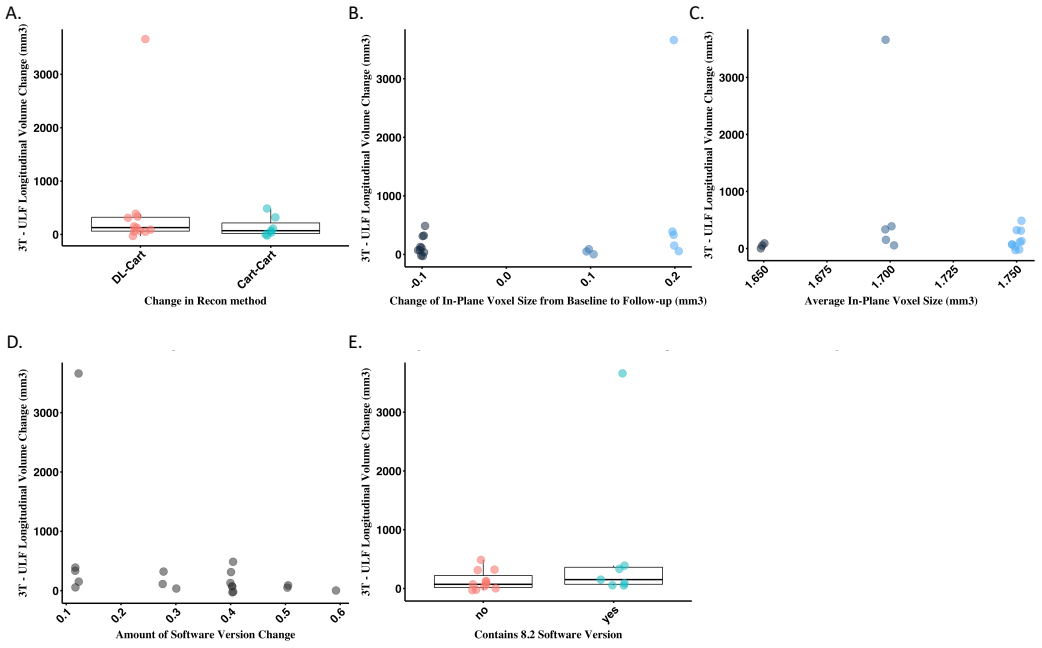


Figure D.9: Difference between 3T and ULF measured longitudinal volume change for NLs by ULF acquisition-related metrics. Difference between 3T and ULF measured longitudinal volume change by: A) change in reconstruction method, B) voxel size change, C) average voxel size, D) software version change, and E) presence of version 8.2.

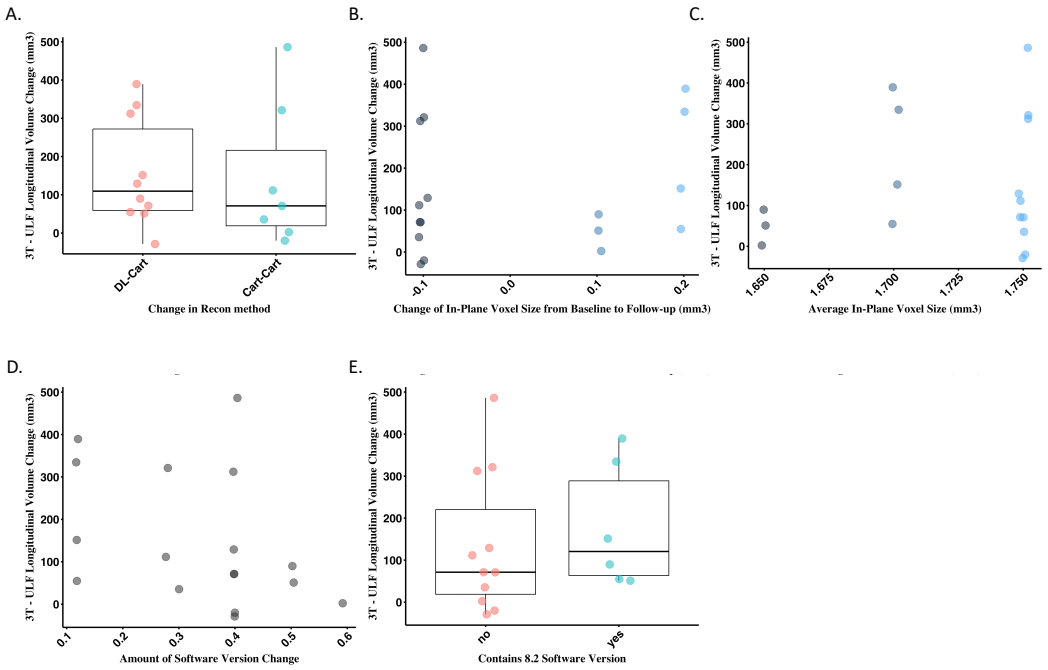


Figure D.10: Difference between 3T and ULF measured longitudinal volume change for NLs by ULF acquisition-related metrics, after removing outlier from figure D.9. Difference between 3T and ULF measured longitudinal volume change by: A) change in reconstruction method, B) voxel size change, C) average voxel size, D) software version change, and E) presence of version 8.2.

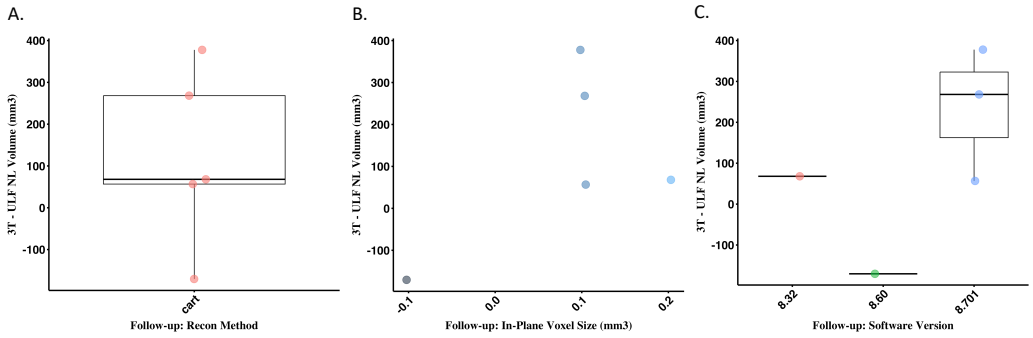


Figure D.11: Difference between 3T and ULF NL volume by ULF follow-up image acquisition parameters. Difference between 3T and ULF NL volume by follow-up image: A) reconstruction method B) in-plane voxel size, C) software version

D.3 Supplementary data for Chapter 4.3

D.3.1 Effect of FouSR Sampling Size on Quantitative Evaluation Metrics

To determine the amount of information to recover in the under-sampled direction, the ratio of the through-slice thickness (5 mm) to the in-plane reso-

lution (1.7 mm) was considered, resulting in a 2.94-fold increase in resolution compared to the original through-plane resolution. Based on this calculation, approximately 54 of the 160 frequency-space lines in the axial direction would theoretically need to be replaced.

In practice, 64 lines were replaced. This choice was guided by stepwise testing of different matrix sizes (0–100) and comparison of resulting quantitative metrics (SNR, CNR, lesion conspicuity, and image sharpness, Figure D.12). These tests showed that slightly oversampling beyond the theoretical minimum improved lesion conspicuity and CNR without substantially affecting SNR (Figure D.12, likely because the outer high-frequency components contribute more edge information than noise.

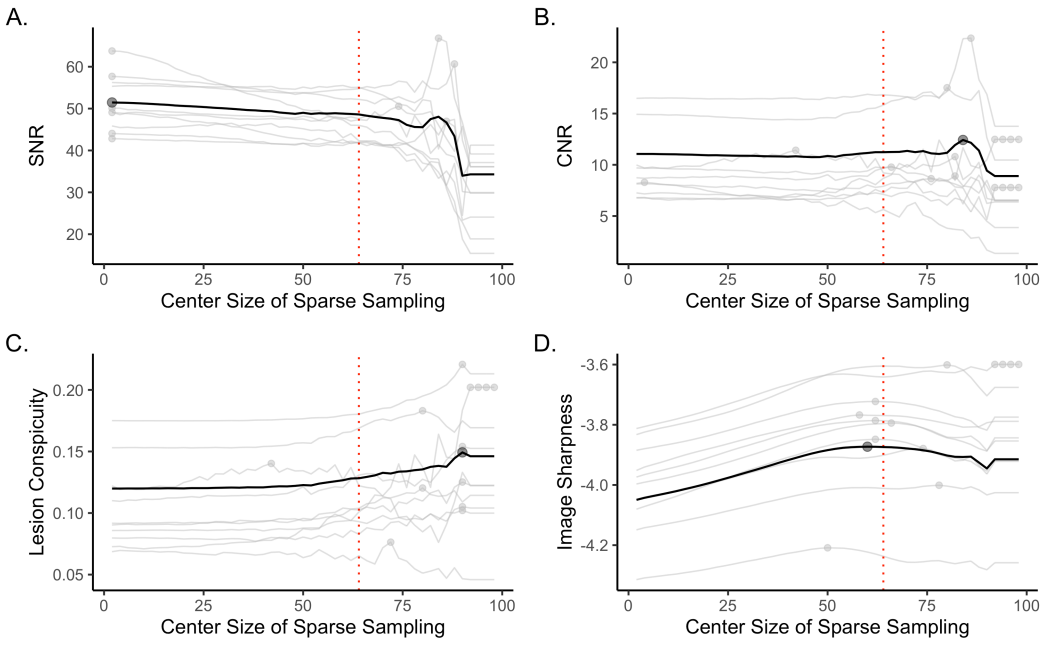


Figure D.12: The effect of varying the centre matrix size in FouSR sparse sampling on quantitative image metrics. For the 10 in-vivo participants, Individual (grey) and group average (black) lines show changes in A) SNR, B) CNR, C) Lesion Conspicuity, and D) Image Sharpness across different centre sizes of FouSR. Points indicate the centre size which results in the maximum image quality for each participant. The dotted red line marks the centre size of 64, which was used in the FouSR algorithm.

Appendix E

Supplementary data for Chapter 5.4

This appendix includes supplementary analyses and detailed statistical outputs supporting the findings reported in Chapter 5.4.

E.1 Extended Cross-Sectional Analyses to All Automatic ULF Segmentation Algorithms

In the main text, Section 5.3.1 shows the results of whether ULF-derived segmentation volumes can capture clinically relevant correlations observed with 3T reference segmentations. In Chapter 4.1, a range of automatic ULF WML segmentation algorithms were assessed against both ULF reference segmentations (ULF Aided) and 3T reference segmentations. In that analysis, the PLAn models demonstrated the best overall segmentation performance. Adding a T1-weighted contrast to ULF FLAIR did not meaningfully improve PLAn segmentation accuracy. Based on these results, Section 5.3.1 focused on evaluating whether ULF segmentations (using ULF Aided and PLAn 2 FLAIR) can reproduce clinically meaningful relationships observed with 3T reference volumes. This appendix extends these analyses by including correlations with all segmentation algorithms evaluated in Chapter 4.1.

Figure E.1 shows box plots of total lesion volume across radiologically defined lesion burden categories for each segmentation method. Corresponding pairwise Kruskal-Wallis comparisons (adjusted for multiple comparisons) are reported in Table E.1. All ULF segmentation methods were able to capture significant volume differences between lesion burden categories 1 vs. 2 and 1 vs. 3, consistent with 3T reference segmentations, except for MiMoSA. This

aligns with findings from Chapter 4.1, where MiMoSA underperformed in both qualitative and quantitative evaluations.

Notably, SynthSeg, nnU-Net FLAIR/T1w, and PLAn 1 FLAIR/T1w segmentation volumes showed a significant difference between lesion burden categories 2 and 3, which was not observed in any other segmentation method, including the reference segmentations. This may reflect a tendency for SynthSeg to overestimate lesion volume, particularly in participants with higher lesion burden, and for nnU-Net FLAIR to underestimate volume in those with lower burden, potentially exaggerating subtle differences between intermediate lesion burden groups. Given that PLAn 1 FLAIR/T1w is closely related to other PLAn models that did not show this effect, the observed significance for PLAn 1 FLAIR/T1w is likely driven by unequal group sizes and the relatively small number of participants in lesion burden category 3, where even minor differences in volume between segmentation methods can reach statistical significance.

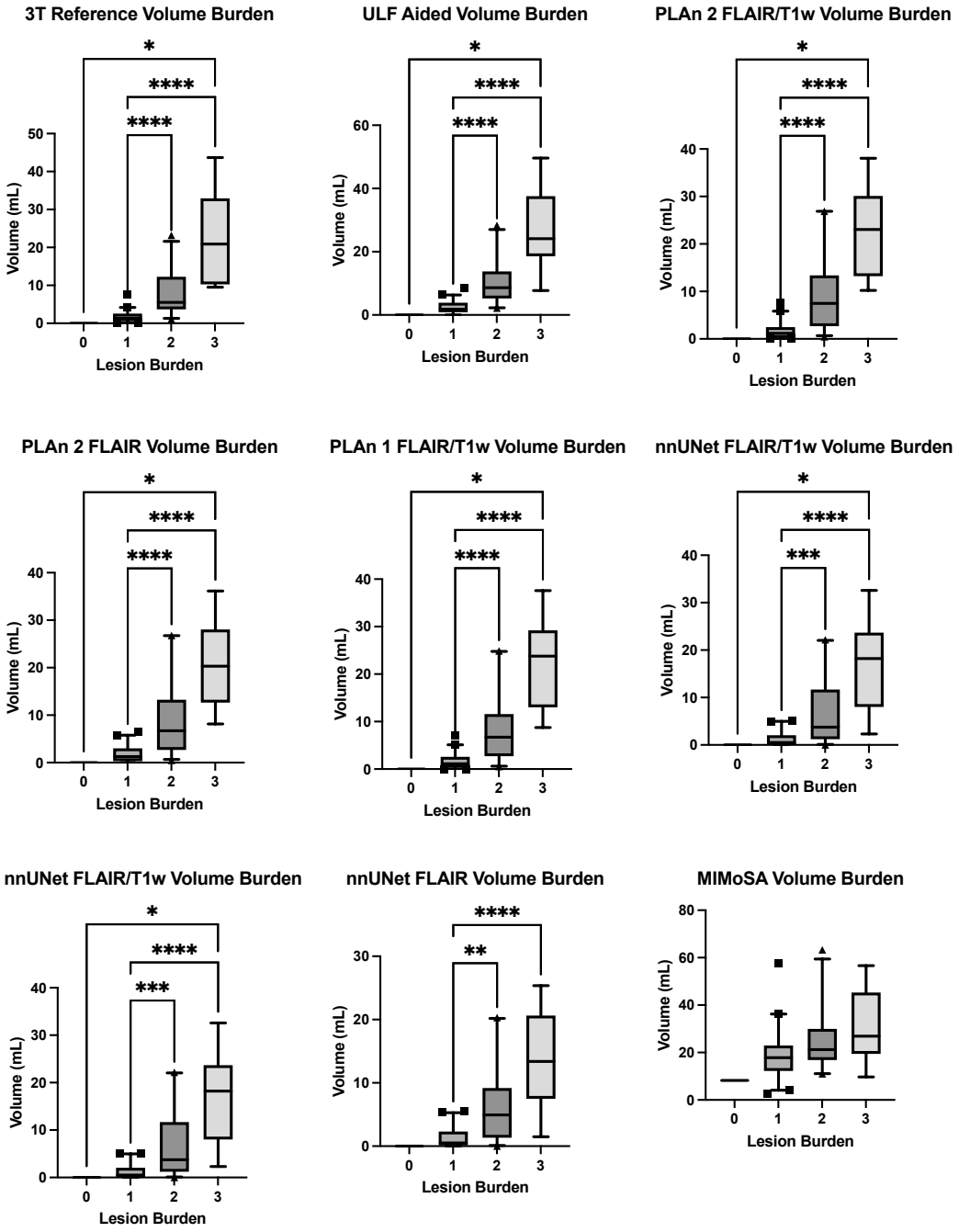


Figure E.1: Total Lesion Volume calculated using 3T and ULF Reference Segmentations and Automatic ULF Segmentation Algorithm Outputs by Radiological Lesion Burden.

Table E.1: P-Values from Kruskal-Wallis Dunn's post hoc test results for total lesion volume across radiologically defined lesion burden categories for each WML segmentation method. Including those from Chapter 5.4, Figure 5.1

	Pairwise Comparison of WML Burden		
Method	**1 vs 2**	**1 vs 3**	**2 vs 3**
ULF Aided	< 0.001	< 0.001	0.11
MIMoSA	0.23	0.08	> 0.99
SynthSeg	0.004	< 0.001	0.004
nnU-Net FLAIR	< 0.001	< 0.001	0.09
nnU-Net FLAIR/T1w	< 0.001	< 0.001	0.04
PLAn 1 FLAIR	< 0.001	< 0.001	0.07
PLAn 1 FLAIR/T1w	< 0.001	< 0.001	0.05
PLAn 2 FLAIR	< 0.001	< 0.001	0.09
PLAn 2 FLAIR/T1w	< 0.001	< 0.001	0.07
3T Reference	< 0.001	< 0.001	0.12

Abbreviations: WML = white matter lesion

All ULF segmentation methods, except MiMoSA, captured the expected clinical correlations between total WML volume and Age, EDSS, and Scripps scores (Figure E.2). These findings were further supported by linear regression models: all segmentation methods showed significant positive associations between total lesion volume and Age, with the exception of MiMoSA and nnU-Net FLAIR (Table E.2). In age-adjusted models, total lesion volume remained significantly associated with EDSS (Table E.3) and Scripps (Table E.4) for all methods except MiMoSA. For the 9HPT (Table E.6), significant associations were found using nnU-Net FLAIR, nnU-Net FLAIR/T1w, and SynthSeg segmentations. However, these relationships were not captured by manual reference masks, which may reflect methodological biases. nnU-Net tends to underestimate lesion volume, while SynthSeg often overestimates. Moreover, the small number of available cases with 9HPT scores ($n = 28$) raises the possibility that these findings reflect statistical noise. No significant associations were observed between lesion volume and PASAT (Table E.7) or SDMT (Table E.5) scores in the linear models, including for ULF Aided segmentations, despite a significant unadjusted Spearman correlation in the latter.

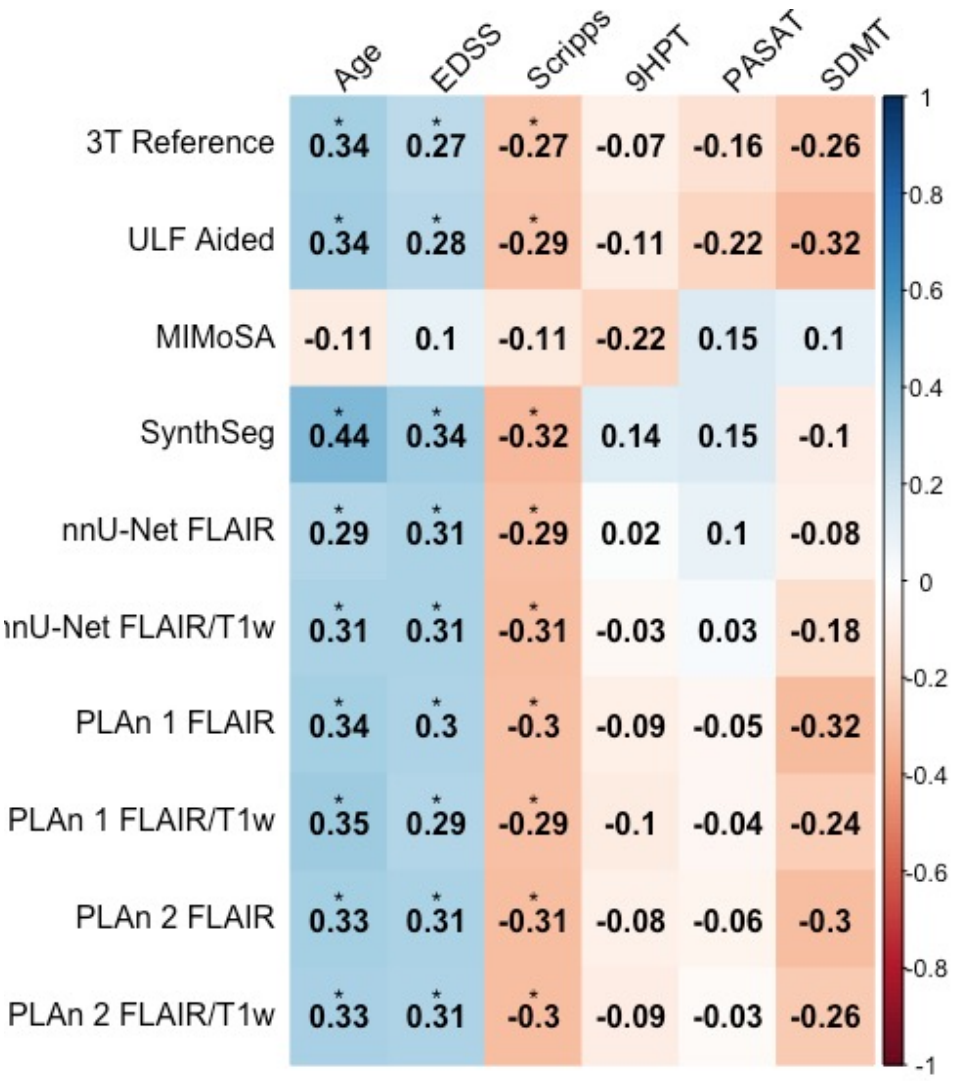


Figure E.2: Spearman correlation matrix showing associations between clinical measures and total lesion volumes derived from different segmentation methods. Clinical variables (age and clinical scores) are shown along the horizontal axis. Segmentation methods are shown on the vertical axis. Asterisks (*) indicate statistically significant correlations ($p < 0.05$). Correlation strengths and directions are colour-coded.

Table E.2: Linear regression results showing associations between age and total lesion volume across segmentation methods. Reported are parameter estimates (β_1), SE, test statistics, p-values, DoF, and R^2 values

Dependent Variable	**Segmentation Method**	**Parameter Est.**	**Variable**	**Est.**	**SE**	**Stat.**	**p-value**	**DoF**	R^2
Age	3T Reference	β_1	Total Lesion Volume	0.38	0.16	2.41	0.02	70.00	0.06
Age	ULF Aided	β_1	Total Lesion Volume	0.35	0.14	2.54	0.01	70.00	0.07
Age	SynthSeg	β_1	Total Lesion Volume	0.94	0.27	3.50	0.00	70.00	0.14
Age	PLAn 2 FLAIR/T1w	β_1	Total Lesion Volume	0.42	0.15	2.73	0.01	70.00	0.08
Age	PLAn 2 FLAIR	β_1	Total Lesion Volume	0.41	0.16	2.49	0.02	70.00	0.07
Age	PLAn 1 FLAIR T1w/FLAIR	β_1	Total Lesion Volume	0.45	0.15	2.89	0.01	70.00	0.09
Age	PLAn 1 FLAIR	β_1	Total Lesion Volume	0.38	0.16	2.44	0.02	70.00	0.06
Age	nnU-Net FLAIR T1w/FLAIR	β_1	Total Lesion Volume	0.40	0.19	2.09	0.04	70.00	0.05
Age	nnU-Net FLAIR	β_1	Total Lesion Volume	0.42	0.23	1.81	0.08	70.00	0.03
Age	MIMoSA	β_1	Total Lesion Volume	-0.13	0.12	-1.03	0.31	70.00	0.00

Abbreviations: ULF = ultra-low field; Est. = estimate; Stat. = Statistic; DoF = degrees of freedom; FLAIR = Fluid Attenuated Inversion Recovery; SE = standard error;

Table E.3: Age-adjusted linear regression results showing associations between age and total lesion volume with EDSS scores across segmentation methods. Reported are parameter estimates (β_1 for Age, β_2 for Total Lesion Volume), SE, test statistics, p-values, DoF, and R^2 values.

Dependent Variable	**Segmentation Method**	**Parameter Est.**	**Variable**	**Est.**	**SE**	**Stat.**	**p-value**	**DoF**	R^2
EDSS	3T Reference	β_1	Age	0.07	0.03	2.58	0.01	67.00	0.17
		β_2	Total Lesion Volume	0.04	0.02	2.22	0.03	67.00	0.17
EDSS	ULF Aided	β_1	Age	0.06	0.02	2.58	0.01	67.00	0.17
		β_2	Total Lesion Volume	0.04	0.02	2.17	0.03	67.00	0.17
EDSS	SynthSeg	β_1	Age	0.14	0.05	3.03	0.00	67.00	0.20
		β_2	Total Lesion Volume	0.03	0.02	1.68	0.10	67.00	0.20
EDSS	PLAn 2 FLAIR/T1w	β_1	Age	0.08	0.02	3.13	0.00	67.00	0.20
		β_2	Total Lesion Volume	0.04	0.02	1.99	0.05	67.00	0.20
EDSS	PLAn 2 FLAIR	β_1	Age	0.08	0.03	3.13	0.00	67.00	0.20
		β_2	Total Lesion Volume	0.04	0.02	2.09	0.04	67.00	0.20
EDSS	PLAn 1 FLAIR T1w/FLAIR	β_1	Age	0.08	0.03	3.05	0.00	67.00	0.20
		β_2	Total Lesion Volume	0.04	0.02	1.93	0.06	67.00	0.20
EDSS	PLAn 1 FLAIR	β_1	Age	0.08	0.02	3.11	0.00	67.00	0.20
		β_2	Total Lesion Volume	0.04	0.02	2.12	0.04	67.00	0.20
EDSS	nnU-Net FLAIR T1w/FLAIR	β_1	Age	0.10	0.03	3.36	0.00	67.00	0.22
		β_2	Total Lesion Volume	0.04	0.02	2.23	0.03	67.00	0.22
EDSS	nnU-Net FLAIR	β_1	Age	0.11	0.04	3.04	0.00	67.00	0.20
		β_2	Total Lesion Volume	0.04	0.02	2.39	0.02	67.00	0.20
EDSS	MIMoSA	β_1	Age	0.02	0.02	0.99	0.32	67.00	0.10
		β_2	Total Lesion Volume	0.06	0.02	3.03	0.00	67.00	0.10

Abbreviations: EDSS = Expanded Disability Status Scale; ULF = ultra-low field; Est. = estimate; Stat. = test statistic; DoF = degrees of freedom; FLAIR = Fluid Attenuated Inversion Recovery; SE = standard error.

Table E.4: Age-adjusted linear regression results showing associations between age and total lesion volume with Scripps scores across segmentation methods. Reported are parameter estimates (β_1 for Age, β_2 for Total Lesion Volume), SE, test statistics, p-values, DoF, and R^2 values.

Dependent Variable	Segmentation Method	Parameter Est.	Variable	Est.	SE	Stat.	p-value	DoF	R^2
Scripps	3T Reference	β_1	Age	-0.64	0.19	-3.41	0.00	67.00	0.18
		β_2	Total Lesion Volume	-0.17	0.13	-1.25	0.22	67.00	0.18
Scripps	ULF Aided	β_1	Age	-0.62	0.16	-3.87	0.00	67.00	0.21
		β_2	Total Lesion Volume	-0.15	0.13	-1.11	0.27	67.00	0.21
Scripps	SynthSeg	β_1	Age	-1.34	0.34	-3.99	0.00	67.00	0.22
		β_2	Total Lesion Volume	-0.08	0.14	-0.61	0.54	67.00	0.22
Scripps	PLAn 2 FLAIR/T1w	β_1	Age	-0.76	0.18	-4.31	0.00	67.00	0.24
		β_2	Total Lesion Volume	-0.12	0.13	-0.92	0.36	67.00	0.24
Scripps	PLAn 2 FLAIR	β_1	Age	-0.76	0.19	-4.02	0.00	67.00	0.22
		β_2	Total Lesion Volume	-0.14	0.13	-1.10	0.27	67.00	0.22
Scripps	PLAn 1 FLAIR T1w/FLAIR	β_1	Age	-0.75	0.18	-4.17	0.00	67.00	0.23
		β_2	Total Lesion Volume	-0.11	0.13	-0.87	0.39	67.00	0.23
Scripps	PLAn 1 FLAIR	β_1	Age	-0.74	0.18	-4.19	0.00	67.00	0.23
		β_2	Total Lesion Volume	-0.14	0.13	-1.10	0.28	67.00	0.23
Scripps	nnU-Net FLAIR T1w/FLAIR	β_1	Age	-0.89	0.21	-4.20	0.00	67.00	0.24
		β_2	Total Lesion Volume	-0.16	0.13	-1.28	0.21	67.00	0.24
Scripps	nnU-Net FLAIR	β_1	Age	-0.87	0.26	-3.29	0.00	67.00	0.17
		β_2	Total Lesion Volume	-0.20	0.13	-1.52	0.13	67.00	0.17
Scripps	MIMoSA	β_1	Age	-0.17	0.14	-1.21	0.23	67.00	0.05
		β_2	Total Lesion Volume	-0.32	0.14	-2.26	0.03	67.00	0.05

Abbreviations: Scripps = Scripps neurologic rating scale; ULF = ultra-low field; Est. = estimate; Stat. = test statistic; DoF = degrees of freedom; FLAIR = Fluid Attenuated Inversion Recovery; SE = standard error.

Table E.5: Age-adjusted linear regression results showing associations between age and total lesion volume with SDMT scores across segmentation methods. Reported are parameter estimates (β_1 for Age, β_2 for Total Lesion Volume), SE, test statistics, p-values, DoF, and R^2 values.

Dependent Variable	**Segmentation Method**	**Parameter Est.**	**Variable**	**Est.**	**SE**	**Stat.**	**p-value**	**DoF**	R^2
SDMT	3T Reference	β_1	Age	-0.61	0.36	-1.66	0.11	24.00	0.09
		β_2	Total Lesion Volume	-0.32	0.24	-1.30	0.21	24.00	0.09
SDMT	ULF Aided	β_1	Age	-0.35	0.26	-1.32	0.20	24.00	0.06
		β_2	Total Lesion Volume	-0.29	0.25	-1.15	0.26	24.00	0.06
SDMT	SynthSeg	β_1	Age	-1.37	0.66	-2.07	0.05	24.00	0.14
		β_2	Total Lesion Volume	-0.33	0.24	-1.38	0.18	24.00	0.14
SDMT	PLAn 2 FLAIR/T1w	β_1	Age	-0.41	0.32	-1.27	0.21	24.00	0.05
		β_2	Total Lesion Volume	-0.31	0.25	-1.25	0.22	24.00	0.05
SDMT	PLAn 2 FLAIR	β_1	Age	-0.43	0.32	-1.33	0.20	24.00	0.06
		β_2	Total Lesion Volume	-0.31	0.25	-1.26	0.22	24.00	0.06
SDMT	PLAn 1 FLAIR T1w/FLAIR	β_1	Age	-0.43	0.35	-1.22	0.23	24.00	0.05
		β_2	Total Lesion Volume	-0.32	0.25	-1.27	0.22	24.00	0.05
SDMT	PLAn 1 FLAIR	β_1	Age	-0.43	0.32	-1.36	0.19	24.00	0.06
		β_2	Total Lesion Volume	-0.31	0.25	-1.27	0.22	24.00	0.06
SDMT	nnU-Net FLAIR T1w/FLAIR	β_1	Age	-0.51	0.40	-1.28	0.21	24.00	0.05
		β_2	Total Lesion Volume	-0.32	0.25	-1.28	0.21	24.00	0.05
SDMT	nnU-Net FLAIR	β_1	Age	-0.44	0.42	-1.04	0.31	24.00	0.03
		β_2	Total Lesion Volume	-0.32	0.25	-1.28	0.21	24.00	0.03
SDMT	MIMoSA	β_1	Age	-0.13	0.24	-0.53	0.60	24.00	0.00
		β_2	Total Lesion Volume	-0.35	0.26	-1.35	0.19	24.00	0.00

Abbreviations: SDMT = Symbol Digit Modalities Test; ULF = ultra-low field; Est. = estimate; Stat. = test statistic; DoF = degrees of freedom; FLAIR = Fluid Attenuated Inversion Recovery; SE = standard error.

Table E.6: Age-adjusted linear regression results showing associations between age and total lesion volume with 9HPT scores (dominant hand) across segmentation methods. Reported are parameter estimates (β_1 for Age, β_2 for Total Lesion Volume), SE, test statistics, p-values, DoF, and R^2 values.

Dependent Variable	**Segmentation Method**	**Parameter Est.**	**Variable**	**Est.**	**SE**	**Stat.**	**p-value**	**DoF**	R^2
9HPT	3T Reference	β_1	Age	0.57	0.33	1.72	0.10	27.00	0.06
		β_2	Total Lesion Volume	-0.32	0.30	-1.08	0.29	27.00	0.06
9HPT	ULF Aided	β_1	Age	0.48	0.28	1.72	0.10	27.00	0.06
		β_2	Total Lesion Volume	-0.36	0.30	-1.20	0.24	27.00	0.06
9HPT	SynthSeg	β_1	Age	1.96	0.62	3.17	0.00	27.00	0.24
		β_2	Total Lesion Volume	-0.33	0.27	-1.21	0.24	27.00	0.24
9HPT	PLAn 2 FLAIR/T1w	β_1	Age	0.65	0.32	2.05	0.05	27.00	0.10
		β_2	Total Lesion Volume	-0.34	0.29	-1.17	0.25	27.00	0.10
9HPT	PLAn 2 FLAIR	β_1	Age	0.67	0.32	2.07	0.05	27.00	0.10
		β_2	Total Lesion Volume	-0.34	0.29	-1.16	0.26	27.00	0.10
9HPT	PLAn 1 FLAIR T1w/FLAIR	β_1	Age	0.65	0.34	1.93	0.06	27.00	0.09
		β_2	Total Lesion Volume	-0.34	0.30	-1.14	0.27	27.00	0.09
9HPT	PLAn 1 FLAIR	β_1	Age	0.63	0.32	1.99	0.06	27.00	0.09
		β_2	Total Lesion Volume	-0.34	0.29	-1.14	0.26	27.00	0.09
9HPT	nnU-Net FLAIR T1w/FLAIR	β_1	Age	0.79	0.39	2.05	0.05	27.00	0.10
		β_2	Total Lesion Volume	-0.34	0.29	-1.15	0.26	27.00	0.10
9HPT	nnU-Net FLAIR	β_1	Age	0.95	0.45	2.10	0.05	27.00	0.11
		β_2	Total Lesion Volume	-0.33	0.29	-1.11	0.28	27.00	0.11
9HPT	MIMoSA	β_1	Age	0.18	0.26	0.69	0.49	27.00	-0.02
		β_2	Total Lesion Volume	-0.27	0.31	-0.87	0.39	27.00	-0.02

Abbreviations: 9HPT = Nine-Hole Peg Test (Dominant Hand); ULF = ultra-low field; Est. = estimate; Stat. = test statistic; DoF = degrees of freedom; FLAIR = Fluid Attenuated Inversion Recovery; SE = standard error.

Table E.7: Age-adjusted linear regression results showing associations between age and total lesion volume with PASAT scores across segmentation methods. Reported are parameter estimates (β_1 for Age, β_2 for Total Lesion Volume), SE, test statistics, p-values, DoF, and R^2 values.

Dependent Variable	Segmentation Method	Parameter Est.	Variable	Est.	SE	Stat.	p-value	DoF	R^2
PASAT	3T Reference	β_1	Age	-0.17	0.50	-0.35	0.73	20.00	-0.08
		β_2	Total Lesion Volume	-0.10	0.24	-0.41	0.69	20.00	-0.08
PASAT	ULF Aided	β_1	Age	-0.19	0.40	-0.47	0.64	20.00	-0.08
		β_2	Total Lesion Volume	-0.09	0.24	-0.37	0.72	20.00	-0.08
PASAT	SynthSeg	β_1	Age	0.06	0.76	0.08	0.94	20.00	-0.09
		β_2	Total Lesion Volume	-0.10	0.24	-0.40	0.69	20.00	-0.09
PASAT	PLAn 2 FLAIR/T1w	β_1	Age	-0.01	0.36	-0.04	0.97	20.00	-0.09
		β_2	Total Lesion Volume	-0.10	0.24	-0.40	0.69	20.00	-0.09
PASAT	PLAn 2 FLAIR	β_1	Age	-0.01	0.37	-0.02	0.98	20.00	-0.09
		β_2	Total Lesion Volume	-0.10	0.24	-0.40	0.69	20.00	-0.09
PASAT	PLAn 1 FLAIR T1w/FLAIR	β_1	Age	-0.01	0.40	-0.02	0.98	20.00	-0.09
		β_2	Total Lesion Volume	-0.10	0.24	-0.40	0.69	20.00	-0.09
PASAT	PLAn 1 FLAIR	β_1	Age	-0.03	0.38	-0.09	0.93	20.00	-0.09
		β_2	Total Lesion Volume	-0.10	0.24	-0.40	0.69	20.00	-0.09
PASAT	nnU-Net FLAIR T1w/FLAIR	β_1	Age	-0.03	0.45	-0.06	0.95	20.00	-0.09
		β_2	Total Lesion Volume	-0.10	0.24	-0.40	0.69	20.00	-0.09
PASAT	nnU-Net FLAIR	β_1	Age	0.05	0.48	0.10	0.92	20.00	-0.09
		β_2	Total Lesion Volume	-0.10	0.24	-0.41	0.69	20.00	-0.09
PASAT	MIMoSA	β_1	Age	0.07	0.26	0.28	0.79	20.00	-0.09
		β_2	Total Lesion Volume	-0.09	0.24	-0.37	0.72	20.00	-0.09

Abbreviations: PASAT = Paced Auditory Serial Addition Test; ULF = ultra-low field; Est. = estimate; Stat. = test statistic; DoF = degrees of freedom; FLAIR = Fluid Attenuated Inversion Recovery; SE = standard error.

References

Absinta, M., Maric, D., Gharagozloo, M., Garton, T., Smith, M. D., Jin, J., Fitzgerald, K. C., Song, A., Liu, P., Lin, J. P., Wu, T., Johnson, K. R., McGavern, D. B., Schafer, D. P., Calabresi, P. A. and Reich, D. S. (2021), 'A lymphocyte-microglia-astrocyte axis in chronic active multiple sclerosis', *Nature* **597**(7878), 709–714.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34497421*

Absinta, M., Sati, P., Fechner, A., Schindler, M. K., Nair, G. and Reich, D. S. (2018), 'Identification of chronic active multiple sclerosis lesions on 3t mri', *AJNR Am J Neuroradiol* **39**(7), 1233–1238.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29724768*

Absinta, M., Sati, P., Masuzzo, F., Nair, G., Sethi, V., Kolb, H., Ohayon, J., Wu, T., Cortese, I. C. M. and Reich, D. S. (2019), 'Association of chronic active multiple sclerosis lesions with disability in vivo', *JAMA Neurol* **76**(12), 1474–1483.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/31403674*

Absinta, M., Sati, P. and Reich, D. S. (2016a), 'Advanced mri and staging of multiple sclerosis lesions', *Nat Rev Neurol* **12**(6), 358–68.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27125632*

Absinta, M., Sati, P., Schindler, M., Leibovitch, E. C., Ohayon, J., Wu, T., Meani, A., Filippi, M., Jacobson, S., Cortese, I. C. and Reich, D. S. (2016), 'Persistent 7-tesla phase rim predicts poor outcome in new multiple sclerosis patient lesions', *J Clin Invest* **126**(7), 2597–609.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27270171*

Afzali, A. M. and Korn, T. (2025), 'The role of the adaptive immune system in the initiation and persistence of multiple sclerosis', *Semin Immunol*

78, 101947.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/40179676*

Airas, L., Bermel, R. A., Chitnis, T., Hartung, H. P., Nakahara, J., Stuve, O., Williams, M. J., Kieseier, B. C. and Wiendl, H. (2024), 'A review of bruton's tyrosine kinase inhibitors in multiple sclerosis', *Ther Adv Neurol Disord* **17**, 17562864241233041.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38638671*

Al Gburi, M., Mazzola, M., Absinta, M., Gaitan, M. I., Reich, D. S., Dundamadappa, S. K. and Hemond, C. C. (2025), 'Paramagnetic rim lesion formation is predicted by the initial gadolinium-enhancing lesion diameter', *Mult Scler* **31**(3), 263–277.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/39819146*

Alessandrino, F., Pichiecchio, A., Mallucci, G., Ghione, E., Romani, A., Bergamaschi, R. and Bastianello, S. (2018), 'Do mri structured reports for multiple sclerosis contain adequate information for clinical decision making?', *AJR Am J Roentgenol* **210**(1), 24–29.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/28952814*

Altaf, A., Shakir, M., Irshad, H. A., Atif, S., Kumari, U., Islam, O., Kimberly, W. T., Knopp, E., Truwit, C., Siddiqui, K. and Enam, S. A. (2024), 'Applications, limitations and advancements of ultra-low-field magnetic resonance imaging: A scoping review', *Surg Neurol Int* **15**, 218.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38974534*

Altokhis, A. I., Hibbert, A. M., Allen, C. M., Mougin, O., Alotaibi, A., Lim, S. Y., Constantinescu, C. S., Abdel-Fahim, R. and Evangelou, N. (2022), 'Longitudinal clinical study of patients with iron rim lesions in multiple sclerosis', *Mult Scler* **28**(14), 2202–2211.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36000485*

Alu, A., Lei, H., Han, X., Wei, Y. and Wei, X. (2022), 'Btk inhibitors in the treatment of hematological malignancies and inflammatory diseases: mechanisms and clinical studies', *J Hematol Oncol* **15**(1), 138.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36183125*

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**(3), 1037–44.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/15006671*

Arnold, D. L., Belachew, S., Gafson, A. R., Gaetano, L., Bernasconi, C. and Elliott, C. (2021), 'Slowly expanding lesions are a marker of progressive ms - no', *Mult Scler* **27**(11), 1681–1683.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34474615*

Arnold, T. C., Freeman, C. W., Litt, B. and Stein, J. M. (2023), 'Low-field mri: Clinical promise and challenges', *J Magn Reson Imaging* **57**(1), 25–44.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36120962*

Arnold, T. C., Tu, D., Okar, S. V., Nair, G., By, S., Kawatra, K. D., Robert-Fitzgerald, T. E., Desiderio, L. M., Schindler, M. K., Shinohara, R. T., Reich, D. S. and Stein, J. M. (2022), 'Sensitivity of portable low-field magnetic resonance imaging for multiple sclerosis lesions', *Neuroimage Clin* **35**, 103101.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35792417*

Ashburner, J. and Friston, K. J. (2005), 'Unified segmentation', *Neuroimage* **26**(3), 839–51.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/15955494*

Avants, B. B., Epstein, C. L., Grossman, M. and Gee, J. C. (2008), 'Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain', *Med Image Anal* **12**(1), 26–41.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/17659998*

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A. and Gee, J. C. (2011), 'A reproducible evaluation of ants similarity metric performance in brain image registration', *Neuroimage* **54**(3), 2033–44.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/20851191*

Ayde, R., Senft, T., Salameh, N. and Sarracanie, M. (2022), 'Deep learning for fast low-field mri acquisitions', *Sci Rep* **12**(1), 11394.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35794175*

Badovinac, V., Mostarica-Stojkovic, M., Dinarello, C. A. and Stosic-Grujicic, S. (1998), 'Interleukin-1 receptor antagonist suppresses experimental autoimmune encephalomyelitis (eae) in rats by influencing the activation and proliferation of encephalitogenic cells', *J Neuroimmunol* **85**(1), 87–95.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/9627001*

Bagnato, F., Jeffries, N., Richert, N. D., Stone, R. D., Ohayon, J. M., McFarland, H. F. and Frank, J. A. (2003), 'Evolution of t1 black holes in patients with multiple sclerosis imaged monthly for 4 years', *Brain* **126**(Pt 8), 1782–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/12821527*

Bagnato, F., Sati, P., Hemond, C. C., Elliott, C., Gauthier, S. A., Harrison, D. M., Mainero, C., Oh, J., Pitt, D., Shinohara, R. T., Smith, S. A., Trapp, B., Azevedo, C. J., Calabresi, P. A., Henry, R. G., Laule, C., Ontaneda, D., Rooney, W. D., Sicotte, N. L., Reich, D. S. and Absinta, M. (2024), 'Imaging chronic active lesions in multiple sclerosis: a consensus statement', *Brain* .
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38226694*

Bane, O., Hectors, S. J., Wagner, M., Arlinghaus, L. L., Aryal, M. P., Cao, Y., Chenevert, T. L., Fennessy, F., Huang, W., Hylton, N. M., Kalpathy-Cramer, J., Keenan, K. E., Malyarenko, D. I., Mulkern, R. V., Newitt, D. C., Russek, S. E., Stupic, K. F., Tudorica, A., Wilmes, L. J., Yankeelov, T. E., Yen, Y. F., Boss, M. A. and Taouli, B. (2018), 'Accuracy, repeatability, and interplatform reproducibility of t(1) quantification methods used for dce-mri: Results from a multicenter phantom study', *Magn Reson Med* **79**(5), 2564–2575.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/28913930*

Barkhof, F., Filippi, M., van Waesberghe, J. H., Campi, A., Miller, D. H. and Ader, H. J. (1999), 'Interobserver agreement for diagnostic mri criteria in suspected multiple sclerosis', *Neuroradiology* **41**(5), 347–50.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/10379592*

Barquero, G., La Rosa, F., Kebiri, H., Lu, P. J., Rahmanzadeh, R., Weigel, M., Fartaria, M. J., Kober, T., Theaudin, M., Du Pasquier, R., Sati, P., Reich, D. S., Absinta, M., Granziera, C., Maggi, P. and Bach Cuadra, M.

(2020), 'Rimnet: A deep 3d multimodal mri architecture for paramagnetic rim lesion assessment in multiple sclerosis', *Neuroimage Clin* **28**, 102412.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/32961401*

Basser, P. J. (1995), 'Inferring microstructural features and the physiological state of tissues from diffusion-weighted images', *NMR Biomed* **8**(7-8), 333–44.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/8739270*

Bazin, P. L. and Pham, D. L. (2008), 'Homeomorphic brain image segmentation with topological and statistical atlases', *Med Image Anal* **12**(5), 616–25.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/18640069*

Bazin, P. L., Weiss, M., Dinse, J., Schafer, A., Trampel, R. and Turner, R. (2014), 'A computational framework for ultra-high resolution cortical segmentation at 7tesla', *Neuroimage* **93 Pt 2**, 201–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23623972*

Beck, E. S., Mullins, W. A., Dos Santos Silva, J., Filippini, S., Parvathaneni, P., Maranzano, J., Morrison, M., Suto, D. J., Donnay, C., Dieckhaus, H., Luciano, N. J., Sharma, K., Gaitan, M. I., Liu, J., de Zwart, J. A., van Gelderen, P., Cortese, I., Narayanan, S., Duyn, J. H., Nair, G., Sati, P. and Reich, D. S. (2023), 'Cortical lesions uniquely predict motor disability accrual and form rarely in the absence of new white matter lesions in multiple sclerosis', *medRxiv* .
URL: *https://www.ncbi.nlm.nih.gov/pubmed/37886541*

Bermudez, C., Remedios, S. W., Ramadass, K., McHugo, M., Heckers, S., Huo, Y. and Landman, B. A. (2020), 'Generalizing deep whole-brain segmentation for post-contrast mri with transfer learning', *J Med Imaging (Bellingham)* **7**(6), 064004.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33381612*

Bexkens, R., Claessen, F. M., Kodde, I. F., Oh, L. S., Eygendaal, D. and van den Bekerom, M. P. (2018), 'The kappa paradox', *Shoulder Elbow* **10**(4), 308.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30214499*

Billot, B., Greve, D. N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A. V., Iglesias, J. E. and Adni (2023), 'Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining', *Med Image Anal* **86**, 102789.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36857946*

Bland, J. M. and Altman, D. G. (2003), 'Applying the right statistics: analyses of measurement studies', *Ultrasound Obstet Gynecol* **22**(1), 85–93.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/12858311*

Blindenbacher, N., Brunner, E., Asseyer, S., Scheel, M., Siebert, N., Rasche, L., Bellmann-Strobl, J., Brandt, A., Ruprecht, K., Meier, D., Wuerfel, J., Paul, F. and Sinnecker, T. (2020), 'Evaluation of the 'ring sign' and the 'core sign' as a magnetic resonance imaging marker of disease activity and progression in clinically isolated syndrome and early multiple sclerosis', *Mult Scler J Exp Transl Clin* **6**(1), 2055217320915480.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/32284875*

Boffa, G., Razzetta, C., Boccia, D., Garbarino, S., Cipriano, E., Uccelli, A., Lapucci, C., Piana, M. and Inglese, M. (2025), 'Determinants driving the evolution of new multiple sclerosis lesions into chronic active or remyelinated states', *Neuroimage Clin* **47**, 103823.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/40527078*

Bonacchi, R., Filippi, M. and Rocca, M. A. (2022), 'Role of artificial intelligence in ms clinical practice', *Neuroimage Clin* **35**, 103065.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35661470*

Bonnier, G., Fischi-Gomez, E., Roche, A., Hilbert, T., Kober, T., Krueger, G. and Granziera, C. (2019), 'Personalized pathology maps to quantify diffuse and focal brain damage', *Neuroimage Clin* **21**, 101607.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30502080*

Bozsik, B., Toth, E., Polyak, I., Kerekes, F., Szabo, N., Bencsik, K., Klivenyi, P. and Kincses, Z. T. (2022), 'Reproducibility of lesion count in various subregions on mri scans in multiple sclerosis', *Front Neurol* **13**, 843377.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35620784*

Brex, P. A., Ciccarelli, O., O'Riordan, J. I., Sailer, M., Thompson, A. J. and Miller, D. H. (2002), 'A longitudinal study of abnormalities on mri and disability from multiple sclerosis', *N Engl J Med* **346**(3), 158–64.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/11796849*

Brex, P. A., Parker, G. J., Leary, S. M., Molyneux, P. D., Barker, G. J., Davie, C. A., Thompson, A. J. and Miller, D. H. (2000), 'Lesion heterogeneity in multiple sclerosis: a study of the relations between appearances on t1 weighted images, t1 relaxation times, and metabolite concentrations', *J Neurol Neurosurg Psychiatry* **68**(5), 627–32.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/10766895*

Broche, L. M., Ross, P. J., Davies, G. R., MacLeod, M. J. and Lurie, D. J. (2019), 'A whole-body fast field-cycling scanner for clinical molecular imaging studies', *Sci Rep* **9**(1), 10402.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/31320653*

Bruck, W., Bitsch, A., Kolenda, H., Bruck, Y., Stiefel, M. and Lassmann, H. (1997), 'Inflammatory central nervous system demyelination: correlation of magnetic resonance imaging findings with lesion pathology', *Ann Neurol* **42**(5), 783–93.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/9392578*

Bruttger, J., Karram, K., Wortge, S., Regen, T., Marini, F., Hoppmann, N., Klein, M., Blank, T., Yona, S., Wolf, Y., Mack, M., Pinteaux, E., Muller, W., Zipp, F., Binder, H., Bopp, T., Prinz, M., Jung, S. and Waisman, A. (2015), 'Genetic cell ablation reveals clusters of local self-renewing microglia in the mammalian central nervous system', *Immunity* **43**(1), 92–106.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26163371*

Burger, D., Molnarfi, N., Weber, M. S., Brandt, K. J., Benkhoucha, M., Gruaz, L., Chofflon, M., Zamvil, S. S. and Lalive, P. H. (2009), 'Glatiramer acetate increases il-1 receptor antagonist but decreases t cell-induced il-1beta in human monocytes and multiple sclerosis', *Proc Natl Acad Sci U S A* **106**(11), 4355–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/19255448*

Burm, S. M., Peferoen, L. A., Zuiderwijk-Sick, E. A., Haanstra, K. G., t Hart, B. A., van der Valk, P., Amor, S., Bauer, J. and Bajramovic, J. J. (2016),

'Expression of il-1beta in rhesus eae and ms lesions is mainly induced in the cns itself', *J Neuroinflammation* **13**(1), 138.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27266875*

Cadavid, D., Balcer, L., Galetta, S., Aktas, O., Ziemssen, T., Vanopdenbosch, L., Frederiksen, J., Skeen, M., Jaffe, G. J., Butzkueven, H., Ziemssen, F., Massacesi, L., Chai, Y., Xu, L., Freeman, S. and Investigators, R. S. (2017), 'Safety and efficacy of opicinumab in acute optic neuritis (renew): a randomised, placebo-controlled, phase 2 trial', *Lancet Neurol* **16**(3), 189–199.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/28229892*

Cairns, J., Vavasour, I. M., Traboulsee, A., Carruthers, R., Kolind, S. H., Li, D. K. B., Moore, G. R. W. and Laule, C. (2022), 'Diffusely abnormal white matter in multiple sclerosis', *J Neuroimaging* **32**(1), 5–16.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34752664*

Calvi, A., Clarke, M. A., Prados, F., Chard, D., Ciccarelli, O., Alberich, M., Pareto, D., Rodriguez Barranco, M., Sastre-Garriga, J., Tur, C., Rovira, A. and Barkhof, F. (2023), 'Relationship between paramagnetic rim lesions and slowly expanding lesions in multiple sclerosis', *Mult Scler* **29**(3), 352–362.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36515487*

Calvi, A., Haider, L., Prados, F., Tur, C., Chard, D. and Barkhof, F. (2022), 'In vivo imaging of chronic active lesions in multiple sclerosis', *Mult Scler* **28**(5), 683–690.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/32965168*

Cannella, B. and Raine, C. S. (1995), 'The adhesion molecule and cytokine profile of multiple sclerosis lesions', *Ann Neurol* **37**(4), 424–35.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/7536402*

Caranova, M., Soares, J. F., Batista, S., Castelo-Branco, M. and Duarte, J. V. (2023), 'A systematic review of microstructural abnormalities in multiple sclerosis detected with noddi and dti models of diffusion-weighted magnetic resonance imaging', *Magn Reson Imaging* **104**, 61–71.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/37775062*

Cercignani, M. and Wheeler-Kingshott, C. (2019), 'From micro- to macro-structures in multiple sclerosis: what is the added value of diffusion imaging', *NMR Biomed* **32**(4), e3888.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29350435*

Chen, X., Schadelin, S., Lu, P. J., Ocampo-Pineda, M., Weigel, M., Barakovic, M., Ruberte, E., Cagol, A., Marechal, B., Kober, T., Kuhle, J., Kappos, L., Melie-Garcia, L. and Granziera, C. (2023), 'Personalized maps of t1 relaxometry abnormalities provide correlates of disability in multiple sclerosis patients', *Neuroimage Clin* **37**, 103349.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36801600*

Clark, K. A., Manning, A. R., Chen, L., Liu, F., Cao, Q., Bar-Or, A., Shinohara, R. T., Sweeney, E. and Schindler, M. K. (2023), 'Early magnetic resonance imaging features of new paramagnetic rim lesions in multiple sclerosis', *Ann Neurol* **94**(4), 736–744.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/37345334*

Coffman, C. H., White, R., Subramanian, K., Buch, S., Bernitsas, E. and Haacke, E. M. (2022), 'Quantitative susceptibility mapping of both ring and non-ring white matter lesions in relapsing remitting multiple sclerosis', *Magn Reson Imaging* **91**, 45–51.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35605736*

Condon, B., Patterson, J., Jenkins, A., Wyper, D., Hadley, D., Grant, R., Rowan, J. and Teasdale, G. (1987), 'Mr relaxation times of cerebrospinal fluid', *J Comput Assist Tomogr* **11**(2), 203–207. Published Mar-Apr 1987. PMID: 3819116.

Cooley, C. Z., McDaniel, P. C., Stockmann, J. P., Srinivas, S. A., Cauley, S. F., Sliwiak, M., Sappo, C. R., Vaughn, C. F., Guerin, B., Rosen, M. S., Lev, M. H. and Wald, L. L. (2021), 'A portable scanner for magnetic resonance imaging of the brain', *Nat Biomed Eng* **5**(3), 229–239.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33230306*

Cox, R. W. (1996), 'Afni: Software for analysis and visualization of functional magnetic resonance neuroimages', *Computers and Biomedical Research* **29**(3), 162–173.
URL: *¡Go to ISI¿://WOS:A1996UV56700002*

Cree, B. A. C., Hollenbach, J. A., Bove, R., Kirkish, G., Sacco, S., Caverzasi, E., Bischof, A., Gundel, T., Zhu, A. H., Papinutto, N., Stern, W. A., Bevan, C., Romeo, A., Goodin, D. S., Gelfand, J. M., Graves, J., Green, A. J., Wilson, M. R., Zamvil, S. S., Zhao, C., Gomez, R., Ragan, N. R., Rush, G. Q., Barba, P., Santaniello, A., Baranzini, S. E., Oksenberg, J. R., Henry, R. G. and Hauser, S. L. (2019), 'Silent progression in disease activity-free relapsing multiple sclerosis', *Ann Neurol* **85**(5), 653–666.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30851128*

Dal-Bianco, A., Grabner, G., Kronnerwetter, C., Weber, M., Hoftberger, R., Berger, T., Auff, E., Leutmezer, F., Trattnig, S., Lassmann, H., Bagnato, F. and Hametner, S. (2017), 'Slow expansion of multiple sclerosis iron rim lesions: pathology and 7 t magnetic resonance imaging', *Acta Neuropathol* **133**(1), 25–42.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27796537*

Dal-Bianco, A., Grabner, G., Kronnerwetter, C., Weber, M., Kornek, B., Kasprian, G., Berger, T., Leutmezer, F., Rommer, P. S., Trattnig, S., Lassmann, H. and Hametner, S. (2021), 'Long-term evolution of multiple sclerosis iron rim lesions in 7 t mri', *Brain* **144**(3), 833–847.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33484118*

Dal-Bianco, A., Schranzer, R., Grabner, G., Lanzinger, M., Kolbrink, S., Pusswald, G., Altmann, P., Ponleitner, M., Weber, M., Kornek, B., Zebenholzer, K., Schmied, C., Berger, T., Lassmann, H., Trattnig, S., Hametner, S., Leutmezer, F. and Rommer, P. (2021a), 'Iron rims in patients with multiple sclerosis as neurodegenerative marker? a 7-tesla magnetic resonance study', *Front Neurol* **12**, 632749.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34992573*

Danelakis, A., Theoharis, T. and Verganelakis, D. A. (2018), 'Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging', *Comput Med Imaging Graph* **70**, 83–100.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30326367*

de Leeuw den Bouter, M. L., Ippolito, G., O'Reilly, T. P. A., Remis, R. F., van Gijzen, M. B. and Webb, A. G. (2022), 'Deep learning-based single image super-resolution for low-field mr brain images', *Sci Rep* **12**(1), 6362.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35430586*

de Oliveira, M., Piacenti-Silva, M., da Rocha, F. C. G., Santos, J. M., Cardoso, J. D. and Lisboa, P. N. (2022), 'Lesion volume quantification using two convolutional neural networks in mris of multiple sclerosis patients', *Diagnostics* **12**(2).
URL: *¡Go to ISI¿://WOS:000762560400001*

De Rosa, A. P., Benedetto, M., Tagliaferri, S., Bardozzo, F., D'Ambrosio, A., Bisecco, A., Gallo, A., Cirillo, M., Tagliaferri, R. and Esposito, F. (2024), 'Consensus of algorithms for lesion segmentation in brain mri studies of multiple sclerosis', *Sci Rep* **14**(1), 21348.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/39266642*

De Stefano, N., Battaglini, M., Pareto, D., Cortese, R., Zhang, J., Oesingmann, N., Prados, F., Rocca, M. A., Valsasina, P., Vrenken, H., Gandini Wheeler-Kingshott, C. A. M., Filippi, M., Barkhof, F., Rovira, A. and Group, M. S. (2022), 'Magnims recommendations for harmonization of mri data in ms multicenter studies', *Neuroimage Clin* **34**, 102972.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35245791*

Deoni, S. C. L., Bruchhage, M. M. K., Beauchemin, J., Volpe, A., D'Sa, V., Huentelman, M. and Williams, S. C. R. (2021), 'Accessible pediatric neuroimaging using a low field strength mri scanner', *Neuroimage* **238**, 118273.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34146712*

Deoni, S. C. L., Medeiros, P., Deoni, A. T., Burton, P., Beauchemin, J., D'Sa, V., Boskamp, E., By, S., McNulty, C., Mileski, W., Welch, B. E. and Huentelman, M. (2022), 'Development of a mobile low-field mri scanner', *Sci Rep* **12**(1), 5690.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35383255*

Deoni, S. C. L., O'Muircheartaigh, J., Ljungberg, E., Huentelman, M. and Williams, S. C. R. (2022), 'Simultaneous high-resolution t(2) -weighted imaging and quantitative t(2) mapping at low magnetic field strengths using a multiple te and multi-orientation acquisition approach', *Magn Reson Med* **88**(3), 1273–1281.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35553454*

Despotovic, I., Goossens, B. and Philips, W. (2015), 'Mri segmentation of the human brain: challenges, methods, and applications', *Comput Math Methods Med* **2015**, 450341.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25945121*

Dieckhaus, H., Meijboom, R., Okar, S., Wu, T., Parvathaneni, P., Mina, Y., Chandran, S., Waldman, A. D., Reich, D. S. and Nair, G. (2022), 'Logistic regression-based model is more efficient than u-net model for reliable whole brain magnetic resonance imaging segmentation', *Top Magn Reson Imaging* **31**(3), 31–39.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35767314*

Dolgin, E. (2021), 'Btk blockers make headway in multiple sclerosis', *Nat Biotechnol* **39**(1), 3–5.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33432223*

Donnay, C., Dieckhaus, H., Tsagkas, C., Gaitan, M. I., Beck, E. S., Mullins, A., Reich, D. S. and Nair, G. (2023), 'Pseudo-label assisted nnu-net enables automatic segmentation of 7t mri from a single acquisition', *Front Neuroimaging* **2**, 1252261. C.D. and H.D. contributed equally to this work.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38107773*

Donnay, C., Okar, S. V., Tsagkas, C., Gaitan, M. I., Poorman, M., Reich, D. S. and Nair, G. (2024), 'Super resolution using sparse sampling at portable ultra-low field mr', *Front Neurol* **15**, 1330203.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38854960*

Duan, Q., Nair, G., Gudino, N., de Zwart, J. A., van Gelderen, P., Murphy-Boesch, J., Reich, D. S., Duyn, J. H. and Merkle, H. (2015), 'A 7t spine array based on electric dipole transmitters', *Magn Reson Med* **74**(4), 1189–97.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26190585*

Dujmovic, I., Mangano, K., Pekmezovic, T., Quattrocchi, C., Mesaros, S., Stojsavljevic, N., Nicoletti, F. and Drulovic, J. (2009), 'The analysis of il-1 beta and its naturally occurring inhibitors in multiple sclerosis: The elevation of il-1 receptor antagonist and il-1 receptor type ii after steroid therapy', *J Neuroimmunol* **207**(1-2), 101–6.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/19162335*

Dworkin, J. D., Linn, K. A., Oguz, I., Fleishman, G. M., Bakshi, R., Nair, G., Calabresi, P. A., Henry, R. G., Oh, J., Papinutto, N., Pelletier, D., Rooney, W., Stern, W., Sicotte, N. L., Reich, D. S., Shinohara, R. T. and North American Imaging in Multiple Sclerosis, C. (2018), 'An automated statistical technique for counting distinct multiple sclerosis lesions', *AJNR Am J Neuroradiol* **39**(4), 626–633.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29472300*

Ehman, R. L. and Felmlee, J. P. (1989), 'Adaptive technique for high-definition mr imaging of moving structures', *Radiology* **173**(1), 255–63.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/2781017*

Eichinger, P., Schon, S., Pongratz, V., Wiestler, H., Zhang, H., Bussas, M., Hoshi, M. M., Kirschke, J., Berthele, A., Zimmer, C., Hemmer, B., Muhlau, M. and Wiestler, B. (2019), 'Accuracy of unenhanced mri in the detection of new brain lesions in multiple sclerosis', *Radiology* **291**(2), 429–435.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30860448*

Elkady, A. M., Elliott, C., Fetco, D., Araujo, D., Karimaghaloo, Z., Ganzetti, M., Clayton, D., Craveiro, L., Kazlauskaite, A., Narayanan, S., Arnold, D. L. and Rudko, D. A. (2025), 'Longitudinal multiparametric quantitative mri evaluation of acute and chronic multiple sclerosis paramagnetic rim lesions', *J Magn Reson Imaging* **61**(4), 1812–1828.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/39239775*

Elkjaer, M. L., Waede, M. R., Kingo, C., Damsbo, K. and Illes, Z. (2023), 'Expression of bruton s tyrosine kinase in different type of brain lesions of multiple sclerosis patients and during experimental demyelination', *Front Immunol* **14**, 1264128.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38022591*

Elliott, C., Belachew, S., Wolinsky, J. S., Hauser, S., Kappos, L., Barkhof, F., Bernasconi, C., Fecker, J., Model, F., Wei, W. and Arnold, D. (2019), 'Chronic white matter lesion activity predicts clinical progression in primary progressive multiple sclerosis', *Brain* **142**(9), 2787–2799.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/31497864*

Elliott, C., Rudko, D. A., Arnold, D. L., Fetco, D., Elkady, A. M., Araujo, D., Zhu, B., Gafson, A., Tian, Z., Belachew, S., Bradley, D. P. and Fisher, E.

(2023), 'Lesion-level correspondence and longitudinal properties of paramagnetic rim and slowly expanding lesions in multiple sclerosis', *Mult Scler* **29**(6), 680–690.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/37036134*

Elliott, C., Wolinsky, J. S., Hauser, S. L., Kappos, L., Barkhof, F., Bernasconi, C., Wei, W., Belachew, S. and Arnold, D. L. (2019), 'Slowly expanding/evolving lesions as a magnetic resonance imaging marker of chronic active multiple sclerosis lesions', *Mult Scler* **25**(14), 1915–1925.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30566027*

Fartaria, M. J., Sati, P., Todea, A., Radue, E. W., Rahmanzadeh, R., O'Brien, K., Reich, D. S., Bach Cuadra, M., Kober, T. and Granziera, C. (2019), 'Automated detection and segmentation of multiple sclerosis lesions using ultra-high-field mp2rage', *Invest Radiol* **54**(6), 356–364.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30829941*

Fischer, J. S., Rudick, R. A., Cutter, G. R. and Reingold, S. C. (1999), 'The multiple sclerosis functional composite measure (msfc): an integrated approach to ms clinical outcome assessment. national ms society clinical outcomes assessment task force', *Mult Scler* **5**(4), 244–50.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/10467383*

Fisniku, L. K., Brex, P. A., Altmann, D. R., Miszkiel, K. A., Benton, C. E., Lanyon, R., Thompson, A. J. and Miller, D. H. (2008), 'Disability and t2 mri lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis', *Brain* **131**(Pt 3), 808–17.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/18234696*

Fortin, J. P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M. and Shinohara, R. T. (2018), 'Harmonization of cortical thickness measurements across scanners and sites', *Neuroimage* **167**, 104–120.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29155184*

Fortin, J. P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., Shinohara, R. T. and Alzheimer's Disease Neuroimaging, I. (2016), 'Removing inter-subject technical variability in magnetic resonance imaging studies',

Neuroimage **132**, 198–212.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26923370*

Fox, R. J., Bar-Or, A., Traboulsee, A., Oreja-Guevara, C., Giovannoni, G., Vermersch, P., Syed, S., Li, Y., Vargas, W. S., Turner, T. J., Wallstroem, E., Reich, D. S. and Group, H. T. (2025), 'Tolebrutinib in nonrelapsing secondary progressive multiple sclerosis', *N Engl J Med* **392**(19), 1883–1892.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/40202696*

Frank, J. A., Stone, L. A., Smith, M. E., Albert, P. S., Maloni, H. and McFarland, H. F. (1994), 'Serial contrast-enhanced magnetic resonance imaging in patients with early relapsing-remitting multiple sclerosis: implications for treatment trials', *Ann Neurol* **36 Suppl**, S86–90.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/8017894*

Frischer, J. M., Bramow, S., Dal-Bianco, A., Lucchinetti, C. F., Rauschka, H., Schmidbauer, M., Laursen, H., Sorensen, P. S. and Lassmann, H. (2009), 'The relation between inflammation and neurodegeneration in multiple sclerosis brains', *Brain* **132**(Pt 5), 1175–89.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/19339255*

Fujimoto, K., Polimeni, J. R., van der Kouwe, A. J., Reuter, M., Kober, T., Benner, T., Fischl, B. and Wald, L. L. (2014), 'Quantitative comparison of cortical surface reconstructions from mp2rage and multi-echo mprage data at 3 and 7 t', *Neuroimage* **90**, 60–73.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24345388*

Garcia-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), 1–18.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23084503*

Garcia-Merino, A. (2021), 'Bruton's tyrosine kinase inhibitors: A new generation of promising agents for multiple sclerosis therapy', *Cells* **10**(10).
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34685540*

Geethanath, S. and Vaughan, J. T., J. (2019), 'Accessible magnetic resonance imaging: A review', *J Magn Reson Imaging* **49**(7), e65–e77.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30637891*

Geladaris, A., Torke, S. and Weber, M. S. (2022), 'Bruton's tyrosine kinase inhibitors in multiple sclerosis: Pioneering the path towards treatment of progression?', *CNS Drugs* **36**(10), 1019–1030.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36178589*

Gillen, K. M., Mubarak, M., Park, C., Ponath, G., Zhang, S., Dimov, A., Levine-Ritterman, M., Toro, S., Huang, W., Amici, S., Kaunzner, U. W., Gauthier, S. A., Guerau-de Arellano, M., Wang, Y., Nguyen, T. D. and Pitt, D. (2021), 'Qsm is an imaging biomarker for chronic glial activation in multiple sclerosis lesions', *Ann Clin Transl Neurol* **8**(4), 877–886.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33704933*

Goebel, R. (2012), 'Brainvoyager–past, present, future', *Neuroimage* **62**(2), 748–56.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22289803*

Goodin, D. S., Traboulsee, A., Knappertz, V., Reder, A. T., Li, D., Langdon, D., Wolf, C., Beckmann, K., Konieczny, A., Ebers, G. C. and Year Long Term Follow-up Study, I. (2012), 'Relationship between early clinical characteristics and long term disability outcomes: 16 year cohort study (follow-up) of the pivotal interferon beta-1b trial in multiple sclerosis', *J Neurol Neurosurg Psychiatry* **83**(3), 282–7.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22193561*

Granziera, C., Wuerfel, J., Barkhof, F., Calabrese, M., De Stefano, N., Enzinger, C., Evangelou, N., Filippi, M., Geurts, J. J. G., Reich, D. S., Rocca, M. A., Ropele, S., Rovira, A., Sati, P., Toosy, A. T., Vrenken, H., Gandini Wheeler-Kingshott, C. A. M., Kappos, L. and Group, M. S. (2021), 'Quantitative magnetic resonance imaging towards clinical application in multiple sclerosis', *Brain* **144**(5), 1296–1311.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33970206*

Green, A. J., Gelfand, J. M., Cree, B. A., Bevan, C., Boscardin, W. J., Mei, F., Inman, J., Arnow, S., Devereux, M., Abounasr, A., Nobuta, H., Zhu, A., Friessen, M., Gerona, R., von Budingen, H. C., Henry, R. G., Hauser, S. L. and Chan, J. R. (2017), 'Clemastine fumarate as a remyelinating therapy for multiple sclerosis (rebuild): a randomised, controlled, double-blind, crossover trial', *Lancet* **390**(10111), 2481–2489.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29029896*

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U. G., Kuker, W., Battaglini, M., Rothwell, P. M. and Jenkinson, M. (2016), 'Bianca (brain intensity abnormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities', *Neuroimage* **141**, 191–205.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27402600*

Guarda, G., Braun, M., Staehli, F., Tardivel, A., Mattmann, C., Forster, I., Farlik, M., Decker, T., Du Pasquier, R. A., Romero, P. and Tschopp, J. (2011), 'Type i interferon inhibits interleukin-1 production and inflammasome activation', *Immunity* **34**(2), 213–23.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/21349431*

Guo, Z., Long, L., Qiu, W., Lu, T., Zhang, L., Shu, Y., Zhang, K., Fang, L. and Chen, S. (2021), 'The distributional characteristics of multiple sclerosis lesions on quantitative susceptibility mapping and their correlation with clinical severity', *Front Neurol* **12**, 647519.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34305779*

Haast, R. A. M., Ivanov, D. and Uludag, K. (2018), 'The impact of b1+ correction on mp2rage cortical t(1) and apparent cortical thickness at 7t', *Hum Brain Mapp* **39**(6), 2412–2425.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29457319*

Haller, S., Haacke, E. M., Thurnher, M. M. and Barkhof, F. (2021), 'Susceptibility-weighted imaging: Technical essentials and clinical neurologic applications', *Radiology* **299**(1), 3–26.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33620291*

Ham, A. S., Hacker, C. T., Guo, J., Sorby-Adams, A., Kimberly, W. T. and Mateen, F. J. (2024), 'Feasibility and tolerability of portable, low-field brain mri for patients with multiple sclerosis', *Mult Scler Relat Disord* **85**, 105515.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38489947*

Harper, J. G., York, E. N., Meijboom, R., Kampaite, A., Thrippleton, M. J., Kearns, P. K. A., Valdes Hernandez, M. D. C., Chandran, S., Waldman, A. D. and Future, M. S. c. (2023), 'Quantitative t(1) brain mapping in early relapsing-remitting multiple sclerosis: longitudinal changes, lesion

heterogeneity and disability', *Eur Radiol* .
URL: *https://www.ncbi.nlm.nih.gov/pubmed/37943312*

Harrison, D. M., Sati, P., Klawiter, E. C., Narayanan, S., Bagnato, F., Beck, E. S., Barker, P., Calvi, A., Cagol, A., Donadieu, M., Duyn, J., Granziera, C., Henry, R. G., Huang, S. Y., Hoff, M. N., Mainero, C., Ontaneda, D., Reich, D. S., Rudko, D. A., Smith, S. A., Trattnig, S., Zurawski, J., Bakshi, R., Gauthier, S. and Laule, C. (2024), 'The use of 7t mri in multiple sclerosis: review and consensus statement from the north american imaging in multiple sclerosis cooperative', *Brain Commun* **6**(5), fcae359.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/39445084*

Hauser, S. L., Bar-Or, A., Comi, G., Giovannoni, G., Hartung, H. P., Hemmer, B., Lublin, F., Montalban, X., Rammohan, K. W., Selmaj, K., Traboulsee, A., Wolinsky, J. S., Arnold, D. L., Klingelschmitt, G., Masterman, D., Fontoura, P., Belachew, S., Chin, P., Mairon, N., Garren, H., Kappos, L., Opera, I. and Investigators, O. I. C. (2017), 'Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis', *N Engl J Med* **376**(3), 221–234.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/28002679*

Hauser, S. L. and Cree, B. A. C. (2020), 'Treatment of multiple sclerosis: A review', *Am J Med* **133**(12), 1380–1390 e2.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/32682869*

Hemond, C. C. and Bakshi, R. (2018), 'Magnetic resonance imaging in multiple sclerosis', *Cold Spring Harb Perspect Med* **8**(5).
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29358319*

Hemond, C. C., Reich, D. S. and Dundamadappa, S. K. (2022), 'Paramagnetic rim lesions in multiple sclerosis: Comparison of visualization at 1.5-t and 3-t mri', *AJR Am J Roentgenol* **219**(1), 120–131.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34851712*

Herx, L. M., Rivest, S. and Yong, V. W. (2000), 'Central nervous system-initiated inflammation and neurotrophism in trauma: Il-1 beta is required for the production of ciliary neurotrophic factor', *J Immunol* **165**(4), 2232–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/10925311*

Hesamian, M. H., Jia, W., He, X. and Kennedy, P. (2019), 'Deep learning techniques for medical image segmentation: Achievements and challenges', *J Digit Imaging* **32**(4), 582–596.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/31144149*

Hoopes, A., Mora, J. S., Dalca, A. V., Fischl, B. and Hoffmann, M. (2022), 'Synthstrip: skull-stripping for any brain image', *Neuroimage* **260**, 119474.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35842095*

Horsfield, M. A., Rocca, M. A., Pagani, E., Storelli, L., Preziosa, P., Messina, R., Camesasca, F., Copetti, M. and Filippi, M. (2016), 'Estimating brain lesion volume change in multiple sclerosis by subtraction of magnetic resonance images', *J Neuroimaging* **26**(4), 395–402.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27019077*

Hwang, H., Rehman, H. Z. U. and Lee, S. (2019), '3d u-net for skull stripping in brain mri', *Applied Sciences-Basel* **9**(3).
URL: *Go to ISI://WOS:000459976200209*

Iglesias, J. E., Billot, B., Balbastre, Y., Magdamo, C., Arnold, S. E., Das, S., Edlow, B. L., Alexander, D. C., Golland, P. and Fischl, B. (2023a), 'Synthsr: A public ai tool to turn heterogeneous clinical brain scans into high-resolution t1-weighted images for 3d morphometry', *Sci Adv* **9**(5), eadd3607.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36724222*

Iglesias, J. E., Billot, B., Balbastre, Y., Tabari, A., Conklin, J., Gilberto Gonzalez, R., Alexander, D. C., Golland, P., Edlow, B. L., Fischl, B. and Initiative, A. D. N. (2021), 'Joint super-resolution and synthesis of 1 mm isotropic mp-rage volumes from clinical mri exams with scans of different orientation, resolution and contrast', *Neuroimage* **237**, 118206.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34048902*

Iglesias, J. E., Schleicher, R., Laguna, S., Billot, B., Schaefer, P., McKaig, B., Goldstein, J. N., Sheth, K. N., Rosen, M. S. and Kimberly, W. T. (2023b), 'Quantitative brain morphometry of portable low-field-strength mri using super-resolution machine learning', *Radiology* **306**(3), e220522.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36346311*

Im, J. E., Khalifa, M., Gregory, A. V., Erickson, B. J. and Kline, T. L. (2024), 'A systematic review on the use of registration-based change tracking methods in longitudinal radiological images', *J Imaging Inform Med* .
URL: *https://www.ncbi.nlm.nih.gov/pubmed/39578321*

Inojosa, H., Schriefer, D. and Ziemssen, T. (2020), 'Clinical outcome measures in multiple sclerosis: A review', *Autoimmun Rev* **19**(5), 102512.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/32173519*

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. and Maier-Hein, K. H. (2021), 'nnu-net: a self-configuring method for deep learning-based biomedical image segmentation', *Nat Methods* **18**(2), 203–211.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33288961*

Islam, K. T., Zhong, S., Zakavi, P., Chen, Z., Kavnoudias, H., Farquharson, S., Durbridge, G., Barth, M., McMahon, K. L., Parizel, P. M., Dwyer, A., Egan, G. F., Law, M. and Chen, Z. (2023), 'Improving portable low-field mri image quality through image-to-image translation using paired low- and high-field images', *Sci Rep* **13**(1), 21183.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38040835*

Jacobs, C. A., Baker, P. E., Roux, E. R., Picha, K. S., Toivola, B., Waugh, S. and Kennedy, M. K. (1991), 'Experimental autoimmune encephalomyelitis is exacerbated by il-1 alpha and suppressed by soluble il-1 receptor', *J Immunol* **146**(9), 2983–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/1826702*

Jenkinson, M., Bannister, P., Brady, M. and Smith, S. (2002), 'Improved optimization for the robust and accurate linear registration and motion correction of brain images', *NeuroImage* **17**(2), 825–841.
URL:
https://www.sciencedirect.com/science/article/pii/S1053811902911328

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. and Smith, S. M. (2012), 'Fsl', *Neuroimage* **62**(2), 782–790.
URL: *ISI://WOS:000306390600032*

Jenkinson, M. and Smith, S. (2001), 'A global optimisation method for robust affine registration of brain images', *Med Image Anal* **5**(2), 143–56.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/11516708*

Jonkman, L., Soriano, A., Amor, S., Barkhof, F., van der Valk, P., Vrenken, H. and Geurts, J. (2015), 'Can ms lesion stages be distinguished with mri? a postmortem mri and histopathology study', *Journal of Neurology* **262**(4), 1074–1080.

Jovicich, J., Marizzoni, M., Sala-Llonch, R., Bosch, B., Bartres-Faz, D., Arnold, J., Benninghoff, J., Wiltfang, J., Roccatagliata, L., Nobili, F., Hensch, T., Trankner, A., Schonknecht, P., Leroy, M., Lopes, R., Bordet, R., Chanoine, V., Ranjeva, J. P., Didic, M., Gros-Dagnac, H., Payoux, P., Zoccatelli, G., Alessandrini, F., Beltramello, A., Bargallo, N., Blin, O., Frisoni, G. B. and PharmaCog, C. (2013), 'Brain morphometry reproducibility in multi-center 3t mri studies: a comparison of cross-sectional and longitudinal segmentations', *Neuroimage* **83**, 472–84.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23668971*

Kamma, E., Lasisi, W., Libner, C., Ng, H. S. and Plemel, J. R. (2022), 'Central nervous system macrophages in progressive multiple sclerosis: relationship to neurodegeneration and therapeutics', *J Neuroinflammation* **19**(1), 45.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35144628*

Kappos, L., Bar-Or, A., Cree, B. A. C., Fox, R. J., Giovannoni, G., Gold, R., Vermersch, P., Arnold, D. L., Arnould, S., Scherz, T., Wolf, C., Wallstrom, E., Dahlke, F. and Investigators, E. C. (2018), 'Siponimod versus placebo in secondary progressive multiple sclerosis (expand): a double-blind, randomised, phase 3 study', *Lancet* **391**(10127), 1263–1273.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29576505*

Keaney, J., Gasser, J., Gillet, G., Scholz, D. and Kadiu, I. (2019), 'Inhibition of bruton's tyrosine kinase modulates microglial phagocytosis: Therapeutic implications for alzheimer's disease', *J Neuroimmune Pharmacol* **14**(3), 448–461.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30758770*

Klistorner, A., Wang, C., Yiannikas, C., Parratt, J., Dwyer, M., Barton, J., Graham, S. L., You, Y., Liu, S. and Barnett, M. H. (2018), 'Evidence of progressive tissue loss in the core of chronic ms lesions: A longitudinal dti study', *Neuroimage Clin* **17**, 1028–1035.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29387524*

Klistorner, S., Barnett, M. H., Graham, S. L., Wang, C. and Klistorner, A. (2022), 'The expansion and severity of chronic ms lesions follows a periventricular gradient', *Mult Scler* **28**(10), 1504–1514.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35296170*

Klistorner, S., Barnett, M. H., Yiannikas, C., Barton, J., Parratt, J., You, Y., Graham, S. L. and Klistorner, A. (2021), 'Expansion of chronic lesions is linked to disease progression in relapsing-remitting multiple sclerosis patients', *Mult Scler* **27**(10), 1533–1542.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33215557*

Klotz, L., Smolders, J., Lehto, J., Matilainen, M., Lutje, L., Buchholz, L., Albrecht, S., Walter, C., Varghese, J., Wiendl, H., Nylund, M., Thomas, C., Gardberg, M., van den Bosch, A. M. R., Airas, L., Huitinga, I. and Kuhlmann, T. (2025), 'Broad rim lesions are a new pathological and imaging biomarker for rapid disease progression in multiple sclerosis', *Nat Med* **31**(6), 2016–2026.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/40301560*

Kober, T., Granziera, C., Ribes, D., Browaeys, P., Schluep, M., Meuli, R., Frackowiak, R., Gruetter, R. and Krueger, G. (2012), 'Mp2rage multiple sclerosis magnetic resonance imaging at 3 t', *Invest Radiol* **47**(6), 346–52.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22543966*

Kolb, H., Absinta, M., Beck, E. S., Ha, S. K., Song, Y., Norato, G., Cortese, I., Sati, P., Nair, G. and Reich, D. S. (2021), '7t mri differentiates remyelinated from demyelinated multiple sclerosis lesions', *Ann Neurol* **90**(4), 612–626.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34390015*

Kolb, H., Al-Louzi, O., Beck, E. S., Sati, P., Absinta, M. and Reich, D. S. (2022), 'From pathology to mri and back: Clinically relevant biomarkers of multiple sclerosis lesions', *Neuroimage Clin* **36**, 103194.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36170753*

Kramer, J., Bar-Or, A., Turner, T. J. and Wiendl, H. (2023), 'Bruton tyrosine kinase inhibitors for multiple sclerosis', *Nat Rev Neurol* **19**(5), 289–304.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/37055617*

Kuhlmann, T., Moccia, M., Coetzee, T., Cohen, J. A., Correale, J., Graves, J., Marrie, R. A., Montalban, X., Yong, V. W., Thompson, A. J., Reich, D. S. and International Advisory Committee on Clinical Trials in Multiple, S. (2023), 'Multiple sclerosis progression: time for a new mechanism-driven framework', *Lancet Neurol* **22**(1), 78–88.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36410373*

Kurtzke, J. F. (1983), 'Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (edss)', *Neurology* **33**(11), 1444–52.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/6685237*

Kushibar, K., Valverde, S., Gonzalez-Villa, S., Bernal, J., Cabezas, M., Oliver, A. and Llado, X. (2019), 'Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction', *Sci Rep* **9**(1), 6742.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/31043688*

La Rosa, F., Abdulkadir, A., Fartaria, M. J., Rahmanzadeh, R., Lu, P. J., Galbusera, R., Barakovic, M., Thiran, J. P., Granziera, C. and Cuadra, M. B. (2020), 'Multiple sclerosis cortical and wm lesion segmentation at 3t mri: a deep learning method based on flair and mp2rage', *Neuroimage Clin* **27**, 102335.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/32663798*

La Rosa, F., Beck, E. S., Maranzano, J., Thiran, J. P., Granziera, C., Reich, D. S., Sati, P. and Cuadra, M. B. (2021), 'Automated detection of cortical lesions with single and multi-contrast 7t mri', *Multiple Sclerosis Journal* **27**(2), 458–459.
URL: *¡Go to ISI¿://WOS:000706771301336*

La Rosa, F., Wynen, M., Al-Louzi, O., Beck, E. S., Huelnhagen, T., Maggi, P., Thiran, J. P., Kober, T., Shinohara, R. T., Sati, P., Reich, D. S., Granziera, C., Absinta, M. and Bach Cuadra, M. (2022), 'Cortical lesions, central vein sign, and paramagnetic rim lesions in multiple sclerosis: Emerging machine learning techniques and future avenues', *Neuroimage Clin* **36**, 103205.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36201950*

Langkammer, C., Krebs, N., Goessler, W., Scheurer, E., Ebner, F., Yen, K., Fazekas, F. and Ropele, S. (2010), 'Quantitative mr imaging of brain iron: a postmortem validation study', *Radiology* **257**(2), 455–62.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/20843991*

Langkammer, C., Schweser, F., Krebs, N., Deistung, A., Goessler, W., Scheurer, E., Sommer, K., Reishofer, G., Yen, K., Fazekas, F., Ropele, S. and Reichenbach, J. R. (2012), 'Quantitative susceptibility mapping (qsm) as a means to measure brain iron? a post mortem validation study', *Neuroimage* **62**(3), 1593–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22634862*

Lassmann, H. (2008), 'The pathologic substrate of magnetic resonance alterations in multiple sclerosis', *Neuroimaging Clin N Am* **18**(4), 563–76, ix.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/19068402*

Lassmann, H. (2013), 'Pathology and disease mechanisms in different stages of multiple sclerosis', *J Neurol Sci* **333**(1-2), 1–4.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23735777*

Lassmann, H. (2017), 'Targets of therapy in progressive ms', *Mult Scler* **23**(12), 1593–1599.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29041864*

Lassmann, H. (2018*a*), 'Multiple sclerosis pathology', *Cold Spring Harb Perspect Med* **8**(3).
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29358320*

Lassmann, H. (2018*b*), 'Multiple sclerosis pathology and its reflection by imaging technologies: Introduction', *Brain Pathol* **28**(5), 721–722.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30375112*

Lau, V., Xiao, L., Zhao, Y., Su, S., Ding, Y., Man, C., Wang, X., Tsang, A., Cao, P., Lau, G. K. K., Leung, G. K. K., Leong, A. T. L. and Wu, E. X. (2023), 'Pushing the limits of low-cost ultra-low-field mri by dual-acquisition deep learning 3d superresolution', *Magn Reson Med* **90**(2), 400–416.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/37010491*

Lee, M. A., Smith, S., Palace, J. and Matthews, P. M. (1998), 'Defining multiple sclerosis disease activity using mri t2-weighted difference imaging', *Brain* **121 (Pt 11)**, 2095–102.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/9827769*

Levesque, S. A., Pare, A., Mailhot, B., Bellver-Landete, V., Kebir, H., Lecuyer, M. A., Alvarez, J. I., Prat, A., de Rivero Vaccari, J. P., Keane, R. W. and Lacroix, S. (2016), 'Myeloid cell transmigration across the cns vasculature triggers il-1beta-driven neuroinflammation during autoimmune encephalomyelitis in mice', *J Exp Med* **213**(6), 929–49.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27139491*

Liguori, M., Meier, D. S., Hildenbrand, P., Healy, B. C., Chitnis, T., Baruch, N. F., Khoury, S. J., Weiner, H. L., Bakshi, R., Barkhof, F. and Guttmann, C. R. (2011), 'One year activity on subtraction mri predicts subsequent 4 year activity and progression in multiple sclerosis', *J Neurol Neurosurg Psychiatry* **82**(10), 1125–31.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/21429902*

Lim, T. R., Suthiphosuwan, S., Micieli, J., Vosoughi, R., Schneider, R., Lin, A. W., Chen, Y. A., Muccilli, A., Marriott, J. J., Selchen, D., Mathur, S., Oh, J. and Bharatha, A. (2024), 'Low-field (64 mt) portable mri for rapid point-of-care diagnosis of dissemination in space in patients presenting with optic neuritis', *AJNR Am J Neuroradiol* **45**(11), 1819–1825.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38926091*

Liu, C., Li, W., Tong, K. A., Yeom, K. W. and Kuzminski, S. (2015), 'Susceptibility-weighted imaging and quantitative susceptibility mapping in the brain', *J Magn Reson Imaging* **42**(1), 23–41.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25270052*

Liu, J., Beck, E. S., Filippini, S., van Gelderen, P., de Zwart, J. A., Norato, G., Sati, P., Al-Louzi, O., Kolb, H., Donadieu, M., Morrison, M., Duyn, J. H. and Reich, D. S. (2021), 'Navigator-guided motion and b0 correction of t2*-weighted magnetic resonance imaging improves multiple sclerosis cortical lesion detection', *Invest Radiol* **56**(7), 409–416.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34086012*

Liu, J., van Gelderen, P., de Zwart, J. A. and Duyn, J. H. (2020), 'Reducing motion sensitivity in 3d high-resolution t(2)*-weighted mri by navigator-based motion and nonlinear magnetic field correction', *Neuroimage* **206**, 116332.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/31689535*

Liu, Y., Leong, A. T. L., Zhao, Y., Xiao, L., Mak, H. K. F., Tsang, A. C. O., Lau, G. K. K., Leung, G. K. K. and Wu, E. X. (2021), 'A low-cost and shielding-free ultra-low-field brain mri scanner', *Nat Commun* **12**(1), 7238.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34907181*

Lopalco, G., Rigante, D., Giannini, M., Galeazzi, M., Lapadula, G., Iannone, F. and Cantarini, L. (2016), 'Safety profile of anakinra in the management of rheumatologic, metabolic and autoinflammatory disorders', *Clin Exp Rheumatol* **34**(3), 531–8.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26940286*

Lou, C., Sati, P., Absinta, M., Clark, K., Dworkin, J. D., Valcarcel, A. M., Schindler, M. K., Reich, D. S., Sweeney, E. M. and Shinohara, R. T. (2021), 'Fully automated detection of paramagnetic rims in multiple sclerosis lesions on 3t susceptibility-based mr imaging', *Neuroimage Clin* **32**, 102796.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34644666*

Lublin, F. D., Reingold, S. C., Cohen, J. A., Cutter, G. R., Sorensen, P. S., Thompson, A. J., Wolinsky, J. S., Balcer, L. J., Banwell, B., Barkhof, F., Bebo, B., J., Calabresi, P. A., Clanet, M., Comi, G., Fox, R. J., Freedman, M. S., Goodman, A. D., Inglese, M., Kappos, L., Kieseier, B. C., Lincoln, J. A., Lubetzki, C., Miller, A. E., Montalban, X., O'Connor, P. W., Petkau, J., Pozzilli, C., Rudick, R. A., Sormani, M. P., Stuve, O., Waubant, E. and Polman, C. H. (2014), 'Defining the clinical course of multiple sclerosis: the 2013 revisions', *Neurology* **83**(3), 278–86.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24871874*

Luchetti, S., Fransen, N. L., van Eden, C. G., Ramaglia, V., Mason, M. and Huitinga, I. (2018), 'Progressive multiple sclerosis patients show substantial lesion activity that correlates with clinical disease severity and sex: a retrospective autopsy cohort analysis', *Acta Neuropathol* **135**(4), 511–528.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29441412*

Ma, Y., Zhang, C., Cabezas, M., Song, Y., Tang, Z., Liu, D., Cai, W., Barnett, M. and Wang, C. (2022), 'Multiple sclerosis lesion analysis in brain magnetic resonance images: Techniques and clinical applications', *IEEE Journal of Biomedical and Health Informatics* **26**(6), 2680–2692.

Macias Islas, M. A. and Ciampi, E. (2019), 'Assessment and impact of cognitive impairment in multiple sclerosis: An overview', *Biomedicines* **7**(1).
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30893874*

Maclaren, J., Herbst, M., Speck, O. and Zaitsev, M. (2013), 'Prospective motion correction in brain imaging: a review', *Magn Reson Med* **69**(3), 621–36.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22570274*

Maggi, P., Sati, P., Nair, G., Cortese, I. C. M., Jacobson, S., Smith, B. R., Nath, A., Ohayon, J., van Pesch, V., Perrotta, G., Pot, C., Theaudin, M., Martinelli, V., Scotti, R., Wu, T., Du Pasquier, R., Calabresi, P. A., Filippi, M., Reich, D. S. and Absinta, M. (2020), 'Paramagnetic rim lesions are specific to multiple sclerosis: An international multicenter 3t mri study', *Ann Neurol* **88**(5), 1034–1042.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/32799417*

Magliozzi, R., Howell, O., Vora, A., Serafini, B., Nicholas, R., Puopolo, M., Reynolds, R. and Aloisi, F. (2007), 'Meningeal b-cell follicles in secondary progressive multiple sclerosis associate with early onset of disease and severe cortical pathology', *Brain* **130**(Pt 4), 1089–104.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/17438020*

Maranzano, J., Dadar, M., Zhernovaia, M., Arnold, D. L., Collins, D. L. and Narayanan, S. (2020), 'Automated separation of diffusely abnormal white matter from focal white matter lesions on mri in multiple sclerosis', *Neuroimage* **213**, 116690.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/32119987*

Marques, J. P., Kober, T., Krueger, G., van der Zwaag, W., Van de Moortele, P. F. and Gruetter, R. (2010), 'Mp2rage, a self bias-field corrected sequence for improved segmentation and t1-mapping at high field', *Neuroimage*

49(2), 1271–81.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/19819338*

Martin, D. and Near, S. L. (1995), 'Protective effect of the interleukin-1 receptor antagonist (il-1ra) on experimental allergic encephalomyelitis in rats', *J Neuroimmunol* **61**(2), 241–5.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/7593560*

Martin, E., Aigrot, M. S., Grenningloh, R., Stankoff, B., Lubetzki, C., Boschert, U. and Zalc, B. (2020), 'Bruton's tyrosine kinase inhibition promotes myelin repair', *Brain Plast* **5**(2), 123–133.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33282676*

Martire, M. S., Moiola, L., Rocca, M. A., Filippi, M. and Absinta, M. (2022), 'What is the potential of paramagnetic rim lesions as diagnostic indicators in multiple sclerosis?', *Expert Rev Neurother* **22**(10), 829–837.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36342396*

Matsuki, T., Nakae, S., Sudo, K., Horai, R. and Iwakura, Y. (2006), 'Abnormal t cell activation caused by the imbalance of the il-1/il-1r antagonist system is responsible for the development of experimental autoimmune encephalomyelitis', *Int Immunol* **18**(2), 399–407.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/16415102*

McFarland, H. F., Frank, J. A., Albert, P. S., Smith, M. E., Martin, R., Harris, J. O., Patronas, N., Maloni, H. and McFarlin, D. E. (1992), 'Using gadolinium-enhanced magnetic resonance imaging lesions to monitor disease activity in multiple sclerosis', *Ann Neurol* **32**(6), 758–66.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/1471866*

McGinley, M. P., Goldschmidt, C. H. and Rae-Grant, A. D. (2021), 'Diagnosis and treatment of multiple sclerosis: A review', *JAMA* **325**(8), 765–779.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33620411*

McKinley, R., Wepfer, R., Grunder, L., Aschwanden, F., Fischer, T., Friedli, C., Muri, R., Rummel, C., Verma, R., Weisstanner, C., Wiestler, B., Berger, C., Eichinger, P., Muhlau, M., Reyes, M., Salmen, A., Chan, A., Wiest, R. and Wagner, F. (2020), 'Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with

segmentation confidence', *Neuroimage Clin* **25**, 102104.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/31927500*

Meaton, I., Altokhis, A., Allen, C. M., Clarke, M. A., Sinnecker, T., Meier, D., Enzinger, C., Calabrese, M., De Stefano, N., Pitiot, A., Giorgio, A., Schoonheim, M. M., Paul, F., Pawlak, M. A., Schmidt, R., Granziera, C., Kappos, L., Montalban, X., Rovira, A., Wuerfel, J. and Evangelou, N. (2022), 'Paramagnetic rims are a promising diagnostic imaging biomarker in multiple sclerosis', *Mult Scler* **28**(14), 2212–2220.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36017870*

Mehta, V., Pei, W., Yang, G., Li, S., Swamy, E., Boster, A., Schmalbrock, P. and Pitt, D. (2013), 'Iron is a sensitive biomarker for inflammation in multiple sclerosis lesions', *PLoS One* **8**(3), e57573.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23516409*

Mellergard, J., Edstrom, M., Vrethem, M., Ernerudh, J. and Dahle, C. (2010), 'Natalizumab treatment in multiple sclerosis: marked decline of chemokines and cytokines in cerebrospinal fluid', *Mult Scler* **16**(2), 208–17.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/20007431*

Miller, D. H., Chard, D. T. and Ciccarelli, O. (2012), 'Clinically isolated syndromes', *Lancet Neurol* **11**(2), 157–69.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22265211*

Montalban, X., Hauser, S. L., Kappos, L., Arnold, D. L., Bar-Or, A., Comi, G., de Seze, J., Giovannoni, G., Hartung, H. P., Hemmer, B., Lublin, F., Rammohan, K. W., Selmaj, K., Traboulsee, A., Sauter, A., Masterman, D., Fontoura, P., Belachew, S., Garren, H., Mairon, N., Chin, P., Wolinsky, J. S. and Investigators, O. C. (2017), 'Ocrelizumab versus placebo in primary progressive multiple sclerosis', *N Engl J Med* **376**(3), 209–220.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/28002688*

Moraal, B., Meier, D. S., Poppe, P. A., Geurts, J. J., Vrenken, H., Jonker, W. M., Knol, D. L., van Schijndel, R. A., Pouwels, P. J., Pohl, C., Bauer, L., Sandbrink, R., Guttmann, C. R. and Barkhof, F. (2009), 'Subtraction mr images in a multiple sclerosis multicenter clinical trial setting', *Radiology* **250**(2), 506–14.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/19037018*

Moraal, B., van den Elskamp, I. J., Knol, D. L., Uitdehaag, B. M., Geurts, J. J., Vrenken, H., Pouwels, P. J., van Schijndel, R. A., Meier, D. S., Guttmann, C. R. and Barkhof, F. (2010), 'Long-interval t2-weighted subtraction magnetic resonance imaging: a powerful new outcome measure in multiple sclerosis trials', *Ann Neurol* **67**(5), 667–75.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/20437564*

Mottershead, J. P., Schmierer, K., Clemence, M., Thornton, J. S., Scaravilli, F., Barker, G. J., Tofts, P. S., Newcombe, J., Cuzner, M. L., Ordidge, R. J., McDonald, W. I. and Miller, D. H. (2003), 'High field mri correlates of myelin content and axonal density in multiple sclerosis–a post-mortem study of the spinal cord', *J Neurol* **250**(11), 1293–301.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/14648144*

Mulder, E. R., de Jong, R. A., Knol, D. L., van Schijndel, R. A., Cover, K. S., Visser, P. J., Barkhof, F., Vrenken, H. and Alzheimer's Disease Neuroimaging, I. (2014), 'Hippocampal volume change measurement: quantitative assessment of the reproducibility of expert manual outlining and the automated methods freesurfer and first', *Neuroimage* **92**, 169–81.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24521851*

Ng Kee Kwong, K. C., Mollison, D., Meijboom, R., York, E. N., Kampaite, A., Thrippleton, M. J., Chandran, S. and Waldman, A. D. (2021), 'The prevalence of paramagnetic rim lesions in multiple sclerosis: A systematic review and meta-analysis', *PLoS One* **16**(9), e0256845.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34495999*

Niaz, M. R., Ridwan, A. R., Wu, Y., Alzheimer's Disease Neuroimaging, I., Bennett, D. A. and Arfanakis, K. (2022), 'Development and evaluation of a high resolution 0.5mm isotropic t1-weighted template of the older adult brain', *Neuroimage* **248**, 118869.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34986396*

Oh, J., Arnold, D. L., Cree, B. A. C., Ionete, C., Kim, H. J., Sormani, M. P., Syed, S., Chen, Y., Maxwell, C. R., Benoit, P., Turner, T. J., Wallstroem, E., Wiendl, H., Tolebrutinib Phase, G. and Trial, G. (2025), 'Tolebrutinib versus teriflunomide in relapsing multiple sclerosis', *N Engl J Med*

392(19), 1893–1904.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/40202623*

Okar, S. V., Nair, G., Kawatra, K. D., Thommana, A. A., Donnay, C. A., Gaitan, M. I., Stein, J. M. and Reich, D. S. (2025), 'High-field-blinded assessment of portable ultra-low-field brain mri for multiple sclerosis', *J Neuroimaging* **35**(1), e70005.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/39815369*

Okubo, G., Okada, T., Yamamoto, A., Kanagaki, M., Fushimi, Y., Okada, T., Murata, K. and Togashi, K. (2016), 'Mp2rage for deep gray matter measurement of the brain: A comparative study with mprage', *J Magn Reson Imaging* **43**(1), 55–62.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26032895*

O'Reilly, T., Teeuwisse, W. M. and Webb, A. G. (2019), 'Three-dimensional mri in a homogenous 27 cm diameter bore halbach array magnet', *J Magn Reson* **307**, 106578.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/31470234*

Ouellette, R., Bergendal, A., Shams, S., Martola, J., Mainero, C., Kristoffersen Wiberg, M., Fredrikson, S. and Granberg, T. (2018), 'Lesion accumulation is predictive of long-term cognitive decline in multiple sclerosis', *Mult Scler Relat Disord* **21**, 110–116.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29550717*

Patel, N., Horsfield, M. A., Banahan, C., Thomas, A. G., Nath, M., Nath, J., Ambrosi, P. B. and Chung, E. M. L. (2017), 'Detection of focal longitudinal changes in the brain by subtraction of mr images', *American Journal of Neuroradiology* **38**(5), 923–927.
URL: *¡Go to ISI¿://WOS:000402022000010*

Piehl, F. (2021), 'Current and emerging disease-modulatory therapies and treatment targets for multiple sclerosis', *J Intern Med* **289**(6), 771–791.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33258193*

Popescu, B. F., Pirko, I. and Lucchinetti, C. F. (2013), 'Pathology of multiple sclerosis: where do we stand?', *Continuum (Minneap Minn)* **19**(4 Multiple Sclerosis), 901–21.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23917093*

Portaccio, E., Bellinvia, A., Fonderico, M., Pasto, L., Razzolini, L., Totaro, R., Spitaleri, D., Lugaresi, A., Cocco, E., Onofrj, M., Di Palma, F., Patti, F., Maimone, D., Valentino, P., Confalonieri, P., Protti, A., Sola, P., Lus, G., Maniscalco, G. T., Brescia Morra, V., Salemi, G., Granella, F., Pesci, I., Bergamaschi, R., Aguglia, U., Vianello, M., Simone, M., Lepore, V., Iaffaldano, P., Filippi, M., Trojano, M. and Amato, M. P. (2022), 'Progression is independent of relapse activity in early multiple sclerosis: a real-life cohort study', *Brain* **145**(8), 2796–2805.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35325059*

Preziosa, P., Pagani, E., Meani, A., Moiola, L., Rodegher, M., Filippi, M. and Rocca, M. A. (2022), 'Slowly expanding lesions predict 9-year multiple sclerosis disease progression', *Neurol Neuroimmunol Neuroinflamm* **9**(2).
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35105685*

Prins, M., Eriksson, C., Wierinckx, A., Bol, J. G., Binnekade, R., Tilders, F. J. and Van Dam, A. M. (2013), 'Interleukin-1beta and interleukin-1 receptor antagonist appear in grey matter additionally to white matter lesions during experimental multiple sclerosis', *PLoS One* **8**(12), e83835.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24376764*

Rahmanzadeh, R., Galbusera, R., Lu, P. J., Bahn, E., Weigel, M., Barakovic, M., Franz, J., Nguyen, T. D., Spincemaille, P., Schiavi, S., Daducci, A., La Rosa, F., Absinta, M., Sati, P., Bach Cuadra, M., Radue, E. W., Leppert, D., Kuhle, J., Kappos, L., Bruck, W., Reich, D. S., Stadelmann, C., Wang, Y. and Granziera, C. (2022), 'A new advanced mri biomarker for remyelinated lesions in multiple sclerosis', *Ann Neurol* **92**(3), 486–502.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35713309*

Rahmanzadeh, R., Weigel, M., Lu, P. J., Melie-Garcia, L., Nguyen, T. D., Cagol, A., La Rosa, F., Barakovic, M., Lutti, A., Wang, Y., Bach Cuadra, M., Radue, E. W., Gaetano, L., Kappos, L., Kuhle, J., Magon, S. and Granziera, C. (2022), 'A comparative assessment of myelin-sensitive measures in multiple sclerosis patients and healthy subjects', *Neuroimage Clin* **36**, 103177.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36067611*

Rasoanandrianina, H., Massire, A., Taso, M., Guye, M., Ranjeva, J. P., Kober, T. and Callot, V. (2019), 'Regional t(1) mapping of the whole cervical spinal cord using an optimized mp2rage sequence', *NMR Biomed* **32**(11), e4142.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/31393649*

Reeves, J. A., Bartnik, A., Jakimovski, D., Mohebbi, M., Bergsland, N., Salman, F., Schweser, F., Wilding, G., Weinstock-Guttman, B., Dwyer, M. G. and Zivadinov, R. (2024), 'Associations between paramagnetic rim lesion evolution and clinical and radiologic disease progression in persons with multiple sclerosis', *Neurology* **103**(10), e210004.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/39447104*

Reeves, J. A., Mohebbi, M., Wicks, T., Salman, F., Bartnik, A., Jakimovski, D., Bergsland, N., Schweser, F., Weinstock-Guttman, B., Dwyer, M. G. and Zivadinov, R. (2024), 'Paramagnetic rim lesions predict greater long-term relapse rates and clinical progression over 10 years', *Mult Scler* **30**(4-5), 535–545.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38366920*

Reeves, J. A., Mohebbi, M., Zivadinov, R., Bergsland, N., Dwyer, M. G., Salman, F., Schweser, F. and Jakimovski, D. (2023), 'Reliability of paramagnetic rim lesion classification on quantitative susceptibility mapping (qsm) in people with multiple sclerosis: Single-site experience and systematic review', *Mult Scler Relat Disord* **79**, 104968.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/37716210*

Reeves, J. A., Weinstock, Z., Zivadinov, R., Dwyer, M. G., Bergsland, N., Salman, F., Schweser, F., Weinstock-Guttman, B., Benedict, R. H. and Jakimovski, D. (2023), 'Paramagnetic rim lesions are associated with greater incidence of relapse and worse cognitive recovery following relapse', *Mult Scler* **29**(8), 1033–1038.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/37161349*

Reich, D. S., Arnold, D. L., Vermersch, P., Bar-Or, A., Fox, R. J., Matta, A., Turner, T., Wallstrom, E., Zhang, X., Mares, M., Khabirov, F. A., Traboulsee, A. and Tolebrutinib Phase 2b Study, G. (2021), 'Safety and efficacy of tolebrutinib, an oral brain-penetrant btk inhibitor, in relapsing

multiple sclerosis: a phase 2b, randomised, double-blind, placebo-controlled trial', *Lancet Neurol* **20**(9), 729–738.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34418400*

Reich, D. S., White, R., Cortese, I. C., Vuolo, L., Shea, C. D., Collins, T. L. and Petkau, J. (2015), 'Sample-size calculations for short-term proof-of-concept studies of tissue protection and repair in multiple sclerosis lesions via conventional clinical imaging', *Mult Scler* **21**(13), 1693–704.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25662351*

Ronneberger, O., Fischer, P. and Brox, T. (2015), 'U-net: Convolutional networks for biomedical image segmentation', *Medical Image Computing and Computer-Assisted Intervention, Pt Iii* **9351**, 234–241.
URL: *¡Go to ISI¿://WOS:000365963800028*

Ropele, S., Langkammer, C., Enzinger, C., Fuchs, S. and Fazekas, F. (2011), 'Relaxation time mapping in multiple sclerosis', *Expert Rev Neurother* **11**(3), 441–50.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/21375449*

Rossi, S., Studer, V., Motta, C., Germani, G., Macchiarulo, G., Buttari, F., Mancino, R., Castelli, M., De Chiara, V., Weiss, S., Martino, G., Furlan, R. and Centonze, D. (2014), 'Cerebrospinal fluid detection of interleukin-1beta in phase of remission predicts disease progression in multiple sclerosis', *J Neuroinflammation* **11**, 32.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24548694*

Rovaris, M., Comi, G., Ladkani, D., Wolinsky, J. S., Filippi, M. and European/Canadian Glatiramer Acetate Study, G. (2003), 'Short-term correlations between clinical and mr imaging findings in relapsing-remitting multiple sclerosis', *AJNR Am J Neuroradiol* **24**(1), 75–81.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/12533330*

Sahraian, M. A., Radue, E. W., Haller, S. and Kappos, L. (2010), 'Black holes in multiple sclerosis: definition, evolution, and clinical correlations', *Acta Neurol Scand* **122**(1), 1–8.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/20003089*

Saranathan, M., Tourdias, T., Kerr, A. B., Bernstein, J. D., Kerchner, G. A., Han, M. H. and Rutt, B. K. (2014), 'Optimization of magnetization-prepared 3-dimensional fluid attenuated inversion recovery imaging for lesion detection at 7 t', *Invest Radiol* **49**(5), 290–8.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24566291*

Sarracanie, M., LaPierre, C. D., Salameh, N., Waddington, D. E. J., Witzel, T. and Rosen, M. S. (2015), 'Low-cost high-performance mri', *Sci Rep* **5**, 15177.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26469756*

Schmierer, K., Scaravilli, F., Altmann, D. R., Barker, G. J. and Miller, D. H. (2004), 'Magnetization transfer ratio and myelin in postmortem multiple sclerosis brain', *Ann Neurol* **56**(3), 407–15.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/15349868*

Seewann, A., Kooi, E. J., Roosendaal, S. D., Barkhof, F., van der Valk, P. and Geurts, J. J. (2009), 'Translating pathology in multiple sclerosis: the combination of postmortem imaging, histopathology and clinical findings', *Acta Neurol Scand* **119**(6), 349–55.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/19254283*

Selvaganesan, K., Whitehead, E., DeAlwis, P. M., Schindler, M. K., Inati, S., Saad, Z. S., Ohayon, J. E., Cortese, I. C. M., Smith, B., Steven, J., Nath, A., Reich, D. S., Inati, S. and Nair, G. (2019), 'Robust, atlas-free, automatic segmentation of brain mri in health and disease', *Heliyon* **5**(2), e01226.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/30828660*

Seppi, D., Puthenparampil, M., Federle, L., Ruggero, S., Toffanin, E., Rinaldi, F., Perini, P. and Gallo, P. (2014), 'Cerebrospinal fluid il-1beta correlates with cortical pathology load in multiple sclerosis at clinical onset', *J Neuroimmunol* **270**(1-2), 56–60.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24657029*

Sethi, V., Nair, G., Absinta, M., Sati, P., Venkataraman, A., Ohayon, J., Wu, T., Yang, K., Shea, C., Dewey, B. E., Cortese, I. C. and Reich, D. S. (2017), 'Slowly eroding lesions in multiple sclerosis', *Mult Scler* **23**(3), 464–472.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27339071*

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., Jarso, S., Pham, D. L., Reich, D. S., Crainiceanu, C. M., Life, A. I. B. and Initia, A. D. N. (2014), 'Statistical normalization techniques for magnetic resonance imaging', *Neuroimage-Clinical* **6**, 9–19.
URL: *¡Go to ISI¿://WOS:000349668500002*

Simioni, S., Amaru, F., Bonnier, G., Kober, T., Rotzinger, D., Du Pasquier, R., Schluep, M., Meuli, R., Sbarbati, A., Thiran, J. P., Krueger, G. and Granziera, C. (2014), 'Mp2rage provides new clinically-compatible correlates of mild cognitive deficits in relapsing-remitting multiple sclerosis', *J Neurol* **261**(8), 1606–13.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24912471*

Singer, B. A., Feng, J. and Chiong-Rivero, H. (2024), 'Early use of high-efficacy therapies in multiple sclerosis in the united states: benefits, barriers, and strategies for encouraging adoption', *J Neurol* **271**(6), 3116–3130.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38615277*

Sinnecker, T., Ruberte, E., Schadelin, S., Canova, V., Amann, M., Naegelin, Y., Penner, I. K., Muller, J., Kuhle, J., Decard, B., Derfuss, T., Kappos, L., Granziera, C., Wuerfel, J., Magon, S. and Yaldizli, O. (2020), 'New and enlarging white matter lesions adjacent to the ventricle system and thalamic atrophy are independently associated with lateral ventricular enlargement in multiple sclerosis', *J Neurol* **267**(1), 192–202.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/31612322*

Sipe, J. C., Knobler, R. L., Braheny, S. L., Rice, G. P., Panitch, H. S. and Oldstone, M. B. (1984), 'A neurologic rating scale (nrs) for use in multiple sclerosis', *Neurology* **34**(10), 1368–72.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/6541311*

Smith, S. M., De Stefano, N., Jenkinson, M. and Matthews, P. M. (2001), 'Normalized accurate measurement of longitudinal brain change', *J Comput Assist Tomogr* **25**(3), 466–75.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/11351200*

Smith, S. M., Rao, A., De Stefano, N., Jenkinson, M., Schott, J. M., Matthews, P. M. and Fox, N. C. (2007), 'Longitudinal and cross-sectional

analysis of atrophy in alzheimer's disease: cross-validation of bsi, siena and sienax', *Neuroimage* **36**(4), 1200–6.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/17537648*

Solomon, A. J., Marrie, R. A., Viswanathan, S., Correale, J., Magyari, M., Robertson, N. P., Saylor, D. R., Kaye, W., Rechtman, L., Bae, E., Shinohara, R., King, R., Laurson-Doube, J. and Helme, A. (2023), 'Global barriers to the diagnosis of multiple sclerosis: Data from the multiple sclerosis international federation atlas of ms, third edition', *Neurology* **101**(6), e624–e635.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/37321866*

Spini, M., Choi, S. and Harrison, D. M. (2020), '7t mpflair versus mp2rage for quantifying lesion volume in multiple sclerosis', *J Neuroimaging* **30**(4), 531–536.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/32569408*

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**, 462–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24273728*

Stuber, C., Morawski, M., Schafer, A., Labadie, C., Wahnert, M., Leuze, C., Streicher, M., Barapatre, N., Reimann, K., Geyer, S., Spemann, D. and Turner, R. (2014), 'Myelin and iron concentration in the human brain: a quantitative study of mri contrast', *Neuroimage* **93 Pt 1**, 95–106.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24607447*

Svanera, M., Benini, S., Bontempi, D. and Muckli, L. (2021), 'Cerebrum-7t: Fast and fully volumetric brain segmentation of 7 tesla mr volumes', *Hum Brain Mapp* **42**(17), 5563–5580.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34598307*

Tan, I. L., van Schijndel, R. A., Fazekas, F., Filippi, M., Freitag, P., Miller, D. H., Yousry, T. A., Pouwels, P. J., Ader, H. J. and Barkhof, F. (2002), 'Image registration and subtraction to detect active t(2) lesions in ms: an interobserver study', *J Neurol* **249**(6), 767–73.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/12111312*

Tan, I. L., van Schijndel, R. A., Pouwels, P. J., Ader, H. J. and Barkhof, F. (2002), 'Serial isotropic three-dimensional fast flair imaging: using image registration and subtraction to reveal active multiple sclerosis lesions', *AJR Am J Roentgenol* **179**(3), 777–82.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/12185063*

Taylor, A. J., Salerno, M., Dharmakumar, R. and Jerosch-Herold, M. (2016), 'T1 mapping: Basic techniques and clinical applications', *JACC Cardiovasc Imaging* **9**(1), 67–81.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26762877*

Teipel, S. J., Meindl, T., Grinberg, L., Heinsen, H. and Hampel, H. (2008), 'Novel mri techniques in the assessment of dementia', *Eur J Nucl Med Mol Imaging* **35 Suppl 1**, S58–69.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/18205002*

Thaler, C., Faizy, T. D., Sedlacik, J., Holst, B., Sturner, K., Heesen, C., Stellmann, J. P., Fiehler, J. and Siemonsen, S. (2017), 'T1 recovery is predominantly found in black holes and is associated with clinical improvement in patients with multiple sclerosis', *AJNR Am J Neuroradiol* **38**(2), 264–269.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/28059711*

Thaler, C., Faizy, T., Sedlacik, J., Holst, B., Stellmann, J. P., Young, K. L., Heesen, C., Fiehler, J. and Siemonsen, S. (2015), 'T1- thresholds in black holes increase clinical-radiological correlation in multiple sclerosis patients', *PLoS One* **10**(12), e0144693.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26659852*

Thapa, D., Raahemifar, K., Bobier, W. R. and Lakshminarayanan, V. (2014), 'Comparison of super-resolution algorithms applied to retinal images', *J Biomed Opt* **19**(5), 056002.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24788371*

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M. S., Fujihara, K., Galetta, S. L., Hartung, H. P., Kappos, L., Lublin, F. D., Marrie, R. A., Miller, A. E., Miller, D. H., Montalban, X., Mowry, E. M., Sorensen, P. S., Tintore, M., Traboulsee, A. L., Trojano, M., Uitdehaag, B. M. J., Vukusic,

S., Waubant, E., Weinshenker, B. G., Reingold, S. C. and Cohen, J. A. (2018), 'Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria', *Lancet Neurol* **17**(2), 162–173.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29275977*

Tonietto, M., Poirion, E., Lazzarotto, A., Ricigliano, V., Papeix, C., Bottlaender, M., Bodini, B. and Stankoff, B. (2023), 'Periventricular remyelination failure in multiple sclerosis: a substrate for neurodegeneration', *Brain* **146**(1), 182–194.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/36097347*

Toru Asahina, A., Lu, J., Chugh, P., Sharma, S., Sharma, P., Tan, S., Kovoor, J., Stretton, B., Gupta, A., Sorby-Adams, A., Goh, R., Harroud, A., Clarke, M. A., Evangelou, N., Patel, S., Dwyer, A., Agzarian, M., Bacchi, S. and Slee, M. (2024), 'Prognostic significance of paramagnetic rim lesions in multiple sclerosis: A systematic review', *J Clin Neurosci* **129**, 110810.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/39232367*

Turner, T. J., Brun, P., Gruber, R. C. and Ofengeim, D. (2024), 'Comparative cns pharmacology of the bruton's tyrosine kinase (btk) inhibitor tolebrutinib versus other btk inhibitor candidates for treating multiple sclerosis', *Drugs R D* **24**(2), 263–274.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/38965189*

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y. J., Egan, A., Yushkevich, P. A. and Gee, J. C. (2010), 'N4itk: Improved n3 bias correction', *Ieee Transactions on Medical Imaging* **29**(6), 1310–1320.
URL: *¡Go to ISI¿://WOS:000278535800009*

Uher, T., Vaneckova, M., Sobisek, L., Tyblova, M., Seidl, Z., Krasensky, J., Ramasamy, D., Zivadinov, R., Havrdova, E., Kalincik, T. and Horakova, D. (2017), 'Combining clinical and magnetic resonance imaging markers enhances prediction of 12-year disability in multiple sclerosis', *Mult Scler* **23**(1), 51–61.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27053635*

Vachha, B. and Huang, S. Y. (2021), 'Mri with ultrahigh field strength and high-performance gradients: challenges and opportunities for clinical

neuroimaging at 7 t and beyond', *Eur Radiol Exp* **5**(1), 35.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34435246*

Valcarcel, A. M., Linn, K. A., Vandekar, S. N., Satterthwaite, T. D., Muschelli, J., Calabresi, P. A., Pham, D. L., Martin, M. L. and Shinohara, R. T. (2018), 'Mimosa: An automated method for intermodal segmentation analysis of multiple sclerosis brain lesions', *J Neuroimaging* **28**(4), 389–398.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29516669*

Valverde, J. M., Imani, V., Abdollahzadeh, A., De Feo, R., Prakash, M., Ciszek, R. and Tohka, J. (2021), 'Transfer learning in magnetic resonance brain imaging: A systematic review', *J Imaging* **7**(4).
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34460516*

van der Kouwe, A. J., Benner, T. and Dale, A. M. (2006), 'Real-time rigid body motion correction and shimming using cloverleaf navigators', *Magn Reson Med* **56**(5), 1019–32.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/17029223*

Van Reeth, E., Tham, I. W. K., Tan, C. H. and Poh, C. L. (2012), 'Super-resolution in magnetic resonance imaging: A review', *Concepts in Magnetic Resonance Part A* **40a**(6), 306–325.
URL: *¡Go to ISI¿://WOS:000311302600004*

van Waesberghe, J. H., Kamphorst, W., De Groot, C. J., van Walderveen, M. A., Castelijns, J. A., Ravid, R., Lycklama a Nijeholt, G. J., van der Valk, P., Polman, C. H., Thompson, A. J. and Barkhof, F. (1999), 'Axonal loss in multiple sclerosis lesions: magnetic resonance imaging insights into substrates of disability', *Ann Neurol* **46**(5), 747–54.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/10553992*

Vandewalle, P., Susstrunk, S. and Vetterli, M. (2006), 'A frequency domain approach to registration of aliased images with application to super-resolution', *Eurasip Journal on Applied Signal Processing* .
URL: *¡Go to ISI¿://WOS:000242066700001*

Voelker, M. N., Kraff, O., Brenner, D., Wollrab, A., Weinberger, O., Berger, M. C., Robinson, S., Bogner, W., Wiggins, C., Trampel, R., Stocker, T., Niendorf, T., Quick, H. H., Norris, D. G., Ladd, M. E. and Speck, O.

(2016), 'The traveling heads: multicenter brain imaging at 7 tesla', *MAGMA* **29**(3), 399–415.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27097904*

Voelker, M. N., Kraff, O., Goerke, S., Laun, F. B., Hanspach, J., Pine, K. J., Ehses, P., Zaiss, M., Liebert, A., Straub, S., Eckstein, K., Robinson, S., Nagel, A. N., Stefanescu, M. R., Wollrab, A., Klix, S., Felder, J., Hock, M., Bosch, D., Weiskopf, N., Speck, O., Ladd, M. E. and Quick, H. H. (2021), 'The traveling heads 2.0: Multicenter reproducibility of quantitative imaging methods at 7 tesla', *Neuroimage* **232**, 117910.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33647497*

Wald, L. L., McDaniel, P. C., Witzel, T., Stockmann, J. P. and Cooley, C. Z. (2020), 'Low-cost and portable mri', *J Magn Reson Imaging* **52**(3), 686–696.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/31605435*

Walton, C., King, R., Rechtman, L., Kaye, W., Leray, E., Marrie, R. A., Robertson, N., La Rocca, N., Uitdehaag, B., van der Mei, I., Wallin, M., Helme, A., Angood Napier, C., Rijke, N. and Baneke, P. (2020), 'Rising prevalence of multiple sclerosis worldwide: Insights from the atlas of ms, third edition', *Mult Scler* **26**(14), 1816–1821.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33174475*

Wang, X. D., Li, X. H., Cho, J. W., Russ, B. E., Rajamani, N., Omelchenko, A., Ai, L., Korchmaros, A., Sawiak, S., Benn, R. A., Garcia-Saldivar, P., Wang, Z., Kalin, N. H., Schroeder, C. E., Craddock, R. C., Fox, A. S., Evans, A. C., Messinger, A., Milham, M. P. and Xu, T. (2021), 'U-net model for brain extraction: Trained on humans for transfer to non-human primates', *Neuroimage* **235**.
URL: *¡Go to ISI¿://WOS:000660300400001*

Wattjes, M. P., Ciccarelli, O., Reich, D. S., Banwell, B., de Stefano, N., Enzinger, C., Fazekas, F., Filippi, M., Frederiksen, J., Gasperini, C., Hacohen, Y., Kappos, L., Li, D. K. B., Mankad, K., Montalban, X., Newsome, S. D., Oh, J., Palace, J., Rocca, M. A., Sastre-Garriga, J., Tintoré, M., Traboulsee, A., Vrenken, H., Yousry, T., Barkhof, F., Rovira, A., Multipl, M. R. I., Ctr, C. M. S. and Sclero, N. A. I. M. (2021), '2021 magnims-cmsc-naims consensus recommendations on the use of mri in

patients with multiple sclerosis', *Lancet Neurology* **20**(8), 653–670.
URL: *¡Go to ISI¿://WOS:000677684700018*

Weber, C. E., Wittayer, M., Kraemer, M., Dabringhaus, A., Bail, K., Platten, M., Schirmer, L., Gass, A. and Eisele, P. (2022), 'Long-term dynamics of multiple sclerosis iron rim lesions', *Mult Scler Relat Disord* **57**, 103340.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35158450*

Weinshenker, B. G. (1995), 'The natural history of multiple sclerosis', *Neurol Clin* **13**(1), 119–46.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/7739500*

Weinstock-Guttman, B., Sormani, M. P. and Repovic, P. (2022), 'Predicting long-term disability in multiple sclerosis: A narrative review of current evidence and future directions', *Int J MS Care* **24**(4), 184–188.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35875463*

Welch, E. B., Rossman, P. J., Felmlee, J. P. and Manduca, A. (2004), 'Self-navigated motion correction using moments of spatial projections in radial mri', *Magn Reson Med* **52**(2), 337–45.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/15282816*

Wenzel, N., Wittayer, M., Weber, C. E., Schirmer, L., Platten, M., Gass, A. and Eisele, P. (2022), 'Mri predictors for the conversion from contrast-enhancing to iron rim multiple sclerosis lesions', *J Neurol* **269**(8), 4414–4420.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35332392*

Wisnieff, C., Ramanan, S., Olesik, J., Gauthier, S., Wang, Y. and Pitt, D. (2015), 'Quantitative susceptibility mapping (qsm) of white matter multiple sclerosis lesions: Interpreting positive susceptibility and the presence of iron', *Magn Reson Med* **74**(2), 564–70.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25137340*

Yao, B., Bagnato, F., Matsuura, E., Merkle, H., van Gelderen, P., Cantor, F. K. and Duyn, J. H. (2012), 'Chronic multiple sclerosis lesions: characterization with high-field-strength mr imaging', *Radiology* **262**(1), 206–15.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22084205*

Yao, Y., Nguyen, T. D., Pandya, S., Zhang, Y., Hurtado Rua, S., Kovanlikaya, I., Kuceyeski, A., Liu, Z., Wang, Y. and Gauthier, S. A. (2018), 'Combining quantitative susceptibility mapping with automatic zero reference (qsm0) and myelin water fraction imaging to quantify iron-related myelin damage in chronic active ms lesions', *AJNR Am J Neuroradiol* **39**(2), 303–310.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29242359*

Yoo, D., Choi, Y. A., Rah, C. J., Lee, E., Cai, J., Min, B. J. and Kim, E. H. (2021), 'Signal enhancement of low magnetic field magnetic resonance image using a conventional- and cyclic-generative adversarial network models with unpaired image sets', *Front Oncol* **11**, 660284.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/34046353*

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C. and Gerig, G. (2006), 'User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability', *Neuroimage* **31**(3), 1116–28.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/16545965*

Zeng, C., Gu, L., Liu, Z. and Zhao, S. (2020), 'Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain mri', *Front Neuroinform* **14**, 610967.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/33328949*

Zhang, H., Nguyen, T. D., Zhang, J., Marcille, M., Spincemaille, P., Wang, Y., Gauthier, S. A. and Sweeney, E. M. (2022), 'Qsmrim-net: Imbalance-aware learning for identification of chronic active multiple sclerosis lesions on quantitative susceptibility maps', *Neuroimage Clin* **34**, 102979.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/35247730*

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**(1), 45–57.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/11293691*

Zivadinov, R., Zorzon, M., De Masi, R., Nasuelli, D. and Cazzato, G. (2008), 'Effect of intravenous methylprednisolone on the number, size and confluence of plaques in relapsing-remitting multiple sclerosis', *J Neurol Sci*

267(1-2), 28–35.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/17945260*

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M. C., Kaus, M. R., Haker, S. J., Wells, W. M., Jolesz, F. A. and Kikinis, R. (2004), 'Statistical validation of image segmentation quality based on a spatial overlap index - scientific reports', *Academic Radiology* **11**(2), 178–189.
URL: *Go to ISI://WOS:000188549900009*

Zwanenburg, J. J., Hendrikse, J., Visser, F., Takahara, T. and Luijten, P. R. (2010), 'Fluid attenuated inversion recovery (flair) mri at 7.0 tesla: comparison with 1.5 and 3.0 tesla', *Eur Radiol* **20**(4), 915–22.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/19802613*