Reinforcing the generalizability of spinal cord multiple sclerosis lesion segmentation models

AUTHORS:

Pierre-Louis Benveniste[1], Lisa Eunyoung Lee[2], Alexandre Prat[3], Zachary Vavasour[4], Roger Tam[5], Anthony Traboulsee[6], Shannon Kolind[7], Jiwon Oh[8], Michelle Chen[9], Charidimos Tsagkas[10], Christina Granziera[11], Nilser Laines Medina[12], Mark Muhlau[13], Jan Kirschke[14], Julian McGinnis[15], Daniel S. Reich[16], Christopher Hemond[17], Virginie Callot[18], Sarah Demortière[19], Bertrand Audoin[20], Govind Nair[21], Massimo Filippi[22], Paola Valsasina[23], Maria A Rocca[24], Olga Ciccarelli[25], Marios Yiannakas[26], Tobias Granberg[27], Russell Ouellette[28], Shahamat Tauhid[29], Rohit Bakshi[30], Caterina Mainero[31], Constantina Andrada Treaba[32],

1 NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada; Mila - Quebec AI Institute, Montreal, QC, Canada

2 Department of Medicine (Neurology), University of Toronto, Toronto, ON, Canada; BARLO Multiple Sclerosis Centre & Keenan Research Centre, St. Michael's Hospital, Toronto, ON, Canada

3 Department of neuroscience, Université de Montréal, Montreal, QC, Canada; Neuroimmunology research laboratory, University of Montreal Hospital Research Centre (CRCHUM), Montreal, QC, Canada

4 School of Biomedical Engineering, University of British Columbia, Vancouver, BC, Canada

5 School of Biomedical Engineering, University of British Columbia, Vancouver, BC, Canada

6 Departments of Medicine (Neurology), Physics, Radiology, University of British Columbia, Vancouver, BC, Canada

7 Departments of Medicine (Neurology), Physics, Radiology, University of British Columbia, Vancouver, BC, Canada

8 Department of Medicine (Neurology), University of Toronto, Toronto, ON, Canada; BARLO Multiple Sclerosis Centre & Keenan Research Centre, St. Michael's Hospital, Toronto, ON, Canada

9 NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada

10 Department of Neurology, University Hospital Basel, University of Basel, Basel, Switzerland; National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA

11 Department of Neurology, University Hospital Basel, University of Basel, Basel, Switzerland

12 NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada; Mila - Quebec AI Institute, Montreal, QC, Canada; Aix-Marseille Univ, CNRS, CRMBM, Marseille, France; AP-HM, Hôpital Universitaire Timone, CEMEREM, Marseille, France

13 Department of Neurology, School of Medicine and Health, Technical University of Munich, Munich, Germany; TUM-Neuroimaging Center, School of Medicine and Health, Technical University of Munich, Munich, Germany

14 Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, TUM School of Medicine and Health, Munich, Germany

15 Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, TUM School of Medicine and Health, Munich, Germany; Institute for AI in Medicine, Technical University of Munich, Germany

16 National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA

17 Departments of Neurology, University of Massachusetts Memorial Medical Center and University of Massachusetts Chan Medical School, Worcester, MA, USA;

18 Aix-Marseille Univ, CNRS, CRMBM, Marseille, France; AP-HM, Hôpital Universitaire Timone, CEMEREM, Marseille, France

19 AP-HM, Hôpital Universitaire Timone, Neurology Department, Marseille, France

20 Aix-Marseille Univ, CNRS, CRMBM, Marseille, France; AP-HM, Hôpital Universitaire Timone, CEMEREM, Marseille, France; AP-HM, Hôpital Universitaire Timone, Neurology Department, Marseille, France

21 National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA

22 Neuroimaging Research Unit, Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan, Italy; Neurology Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy; Neurorehabilitation Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy; Neurophysiology Service, IRCCS San Raffaele Scientific Institute, Milan, Italy; Vita-Salute San Raffaele University, Milan, Italy

23 Neuroimaging Research Unit, Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan, Italy

24 Neuroimaging Research Unit, Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan, Italy; Neurology Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy; Vita-Salute San Raffaele University, Milan, Italy

25 Queen Square MS Centre, UCL Institute of Neurology, Faculty of Brain Sciences, University College London, London, UK

26 Queen Square MS Centre, UCL Institute of Neurology, Faculty of Brain Sciences, University College London, London, UK

27 Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden; Department of Neuroradiology, Karolinska University Hospital, Stockholm, Sweden

28 Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden; Department of Neuroradiology, Karolinska University Hospital, Stockholm, Sweden

29 Brigham and Women's Hospital, Harvard Medical School, Boston, USA

30 Brigham and Women's Hospital, Harvard Medical School, Boston, USA

31 Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, USA

32 Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, USA

Anne Kerbrat[33], Elise Bannier[34], Gilles Edan[35], Pierre Labauge[36], Kristin P. O'Grady[37], Seth A Smith[38], Timothy M. Shepherd[39], Erik Charlson[40], Jean-Christophe Brisset[41], Jason Talbott[42], Yaou Liu[43], Hervé Lombaert[44], Julien Cohen-Adad[45]

33 Univ Rennes, Inria, CNRS, Inserm, IRISA UMR 6074, Visages U1128, France; CHU Rennes, Neurology Department, Rennes France

34 Univ Rennes, Inria, CNRS, Inserm, IRISA UMR 6074, Visages U1128, France; CHU Rennes, Radiology Department, Rennes, France

35 Univ Rennes, Inria, CNRS, Inserm, IRISA UMR 6074, Visages U1128, France; CHU Rennes, Neurology Department, Rennes France

36 MS Unit, Department of Neurology, University Hospital of Montpellier, Montpellier, France

37 Vanderbilt University Institute of Imaging Science, Nashville, TN, USA

38 Vanderbilt University Institute of Imaging Science, Nashville, TN, USA

39 Departments of Neurology & Radiology, NYU Langone Medical Center, New York, USA

40 Departments of Neurology & Radiology, NYU Langone Medical Center, New York, USA

41 Brisset JC Ph.D. - Medical Imaging Consulting, Sophia Antipolis, Valbonne, France

42 Department of Radiology and Biomedical Imaging, Zuckerberg San Francisco General Hospital, University of California, San Francisco, CA, USA

43 Department of Radiology, Xuanwu Hospital, Capital Medical University, Beijing 100053, P. R. China; Department of Radiology, Beijing Tiantan Hospital, Capital Medical University, Beijing 100050, P. R. China

44 Mila - Quebec AI Institute, Montreal, QC, Canada; Polytechnique Montréal, Montreal, QC, Canada

45 NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada; Mila - Quebec AI Institute, Montreal, QC, Canada; Functional Neuroimaging Unit, CRIUGM, Université de Montréal, Montreal, QC, Canada; Centre de Recherche du CHU Sainte-Justine, Université de Montréal, Montreal, QC, Canada

Introduction

Spinal cord (SC) imaging has become increasingly central in the diagnosis and monitoring of multiple sclerosis (MS) [1,2]. SC lesions bear strong prognostic significance, with evidence linking their spatial distribution to clinical disability [3–5]. Accurate segmentation of SC lesions is essential for monitoring disease progression. Moreover, despite recent initiatives [6–9], there remains a wide variability in MRI acquisition parameters across institutions. Existing SC lesion segmentation methods lack accessibility [10–12], are typically contrast-specific and often fail to generalise to previously unseen imaging protocols [13–16]. Additionally, inter- and intra-rater variability hinders the precise tracking of lesion changes. Our objective is to develop a robust model for MS lesion segmentation on MRI scans that generalises across different contrasts and imaging parameters. We explore two methods to improve generalizability compared to the current state-of-the-art methods.

Methods

A multi-site dataset (20 sites, 1850 people with MS, 4430 scans) was selected based on the heterogeneity in acquisition parameters and sequences: T1w spin echo (n=23), T2w (n=3061), T2*w (n=548), PSIR (n=363), STIR (n=92), MP2RAGE-UNIT1 (n=343) acquired at 1.5T and 3T on GE, Siemens and Philips MRI systems. The field-of-view coverage varied across sites (brain and upper SC, or SC only), and acquisitions were either 2D (axial: n=2895, sagittal: n=1169) or 3D (n=366), with voxel dimensions ranging from 0.2x0.2x5 mm^3 to 0.8x0.8x9 mm^3. Manual segmentations were collected from expert raters across multiple institutions. We explore the following strategies: (i) **Weighted batch sampling:** In each training batch, images are sampled with probabilities inversely proportional to the square root of the number of samples in each contrast, thereby up-weighting under-represented contrasts [17] ; (ii) **Pretrained model fine-tuning:** We fine-tuned a foundational model, pretrained on over 10,000 CT scans [18], on our multi-contrast dataset. Models were trained under equivalent hyperparameters for fair comparison. Evaluation employed both voxel-wise metrics (Dice coefficient) and lesion-wise metrics (lesion-wise positive predictive value (L-PPV), sensitivity, and F1-score). The results were benchmarked against existing SC lesion segmentation tools available in SpinalCordToolbox (SCT): (a) *sct_deepseg_lesion* for T2w/T2*w [13], (b) *sct_deepseg* for PSIR/STIR [16], and (c) *sct_deepseg* for MP2RAGE-UNIT1 [19].

Results

Both experiments performed better than the baseline model. The average Dice score increased from 0.42 (baseline) to 0.44 with weighted batch sampling (i), and to 0.50 with CT-pretrained model fine-tuning (ii). Fine-tuning yielded the highest performance across most metrics, including Dice, L-PPV and L-F1. Interestingly, weighted sampling yielded slightly higher lesion sensitivity, indicating a trade-off between precision and recall. Contrast-specific analyses revealed strong improvements on under-represented modalities. For PSIR (8% of the dataset), Dice increased from 30.6% (baseline) to 45.8% (ii); for STIR (2% of the dataset), from 27.9% to 59.4% (ii). Even high-frequency contrasts (T2w and T2*w) showed performance gains (+8.8% and +1.3%, respectively) with (ii). Compared to state-of-the-art models, (ii) outperformed (a) and (c) on their respective contrasts. However, it did not surpass (b), which had partial access to the test data during training.

Discussion

Both weighted batch sampling and pretrained model fine-tuning independently improved generalisation, particularly benefiting under-represented contrasts. Our evaluation remains limited as methods (a), (b) and (c) were trained on some of the data used during testing, limiting fair comparisons. Moreover, Dice score, while widely used, is suboptimal for small lesions with uncertain boundaries [20]. Although lesion-wise metrics (L-PPV, L-F1) provide more lesion-centric insight, they rely on binary overlap thresholds and are susceptible to segmentation variability. In [21], we demonstrated that expert neuro-radiologist ratings, using a 1-5 Likert scale, often contradicted voxel-wise metrics: predicted segmentations were sometimes judged to better represent lesion presence than manual annotations, reflecting rater variability [11]. This highlights the need for complementary evaluation frameworks. In [21], we suggested that soft segmentations can improve clinical interpretability and enhance lesion detectability [21]. Nonetheless, expert review remains resource-intensive, emphasising the necessity of developing scalable surrogate evaluation metrics that better correlate with expert review.

Conclusions

Fine-tuning a pretrained CT-based model yielded the best segmentation performance across diverse MRI contrasts, demonstrating the feasibility of cross-modality transfer learning in SC MS lesion segmentation. The model and code will be released as part of SCT, promoting reproducibility and collaborative development.

Figures:

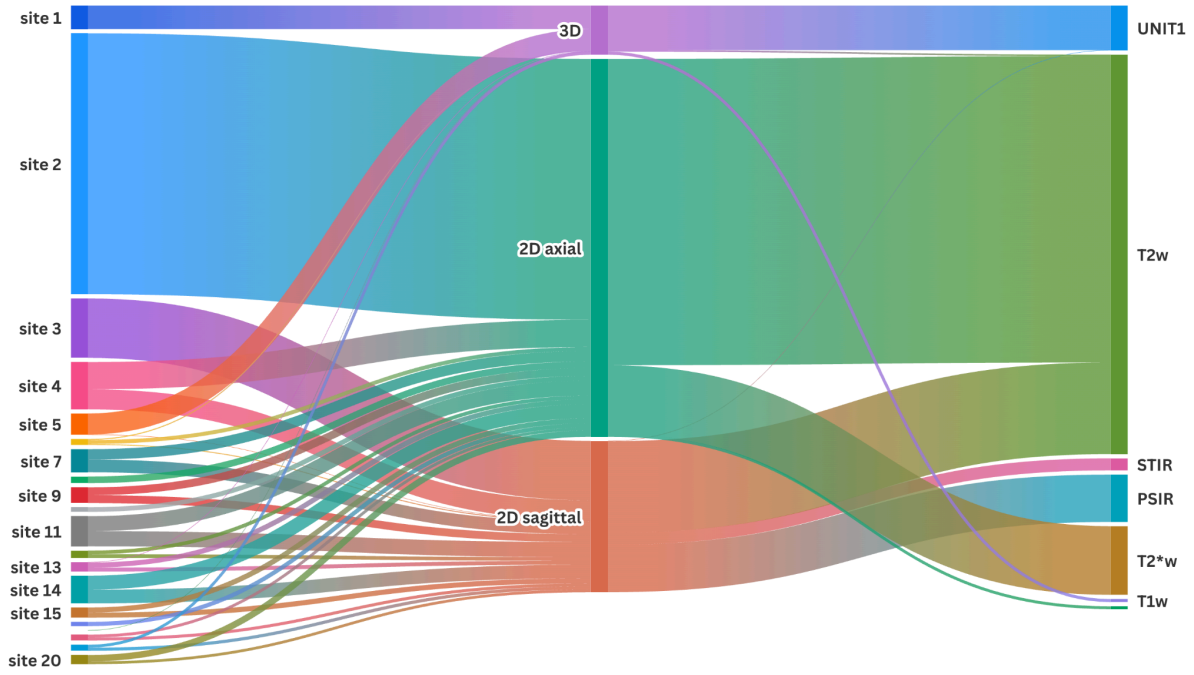

Figure 1: Sankey diagram of scans across MRI datasets. Line thickness corresponds to the number of scans. MRI scan distribution is displayed for acquisition type (3D, 2D sagittal or 2D axial) and MRI contrast.

		Our models			Benchmark		
Method		Baseline model	(i) Sampling method	(ii) Fine-tuning method	(a) sct_deepseg_ lesion (T2w/T2*w)	(b) sct_deepseg (PSIR/STIR)	(c) sct_deepseg (MP2RAGE)
Voxel-wise metrics	Dice	0.42 ± 0.29	0.44 ± 0.30	**0.50 ± 0.33**	0.36 ± 0.37	0.22 ± 0.29	0.21 ± 0.37
Lesion-wise metrics	Recall	0.85 ± 0.26	**0.86 ± 0.25**	0.78 ± 0.33	0.64 ± 0.43	0.68 ± 0.42	0.40 ± 0.48
	PPV	0.62 ± 0.43	0.65 ± 0.41	**0.78 ± 0.36**	0.48 ± 0.46	0.27 ± 0.37	0.24 ± 0.40
	F1-score	0.58 ± 0.39	0.61 ± 0.39	**0.70 ± 0.36**	0.44 ± 0.42	0.26 ± 0.35	0.21 ± 0.38

Figure 2: Comparison of model performance on the test set. Model performance was evaluated on both voxel-wise metrics and lesion-wise metrics. Comparison with state-of-the-art open source models (a, b, c).

	Our models			Benchmark		
	Baseline model	(i) Sampling method	(ii) Fine-tuning method	(a) sct_deepseg_le sion (T2w/T2*w)	(b) sct_deepseg (PSIR/STIR)	(c) sct_deepseg (MP2RAGE)
PSIR (n=34):	0.31 ± 0.23	0.31 ± 0.23	0.46 ± 0.35	0.04 ± 0.17	**0.58 ± 0.31**	0.14 ± 0.21
STIR (n=13):	0.28 ± 0.24	0.50 ± 0.36	0.59 ± 0.37	0.33 ± 0.32	**0.65 ± 0.35**	0.39 ± 0.50
T2*w (n=61):	0.48 ± 0.26	**0.51 ± 0.26**	0.49 ± 0.27	0.42 ± 0.28	0.18 ± 0.28	0.17 ± 0.37
T2w (n=306):	0.41 ± 0.30	0.42 ± 0.31	**0.50 ± 0.35**	0.40 ± 0.38	0.19 ± 0.25	0.20 ± 0.39
UNIT1 (n=35):	0.53 ± 0.26	0.56 ± 0.26	**0.60 ± 0.26**	0.13 ± 0.32	0.00 ± 0.00	0.32 ± 0.24

Figure 3: Comparison of Dice scores per MRI Contrast on the test set. Both methods (i) and (ii) perform well globally, even on under-represented contrasts.

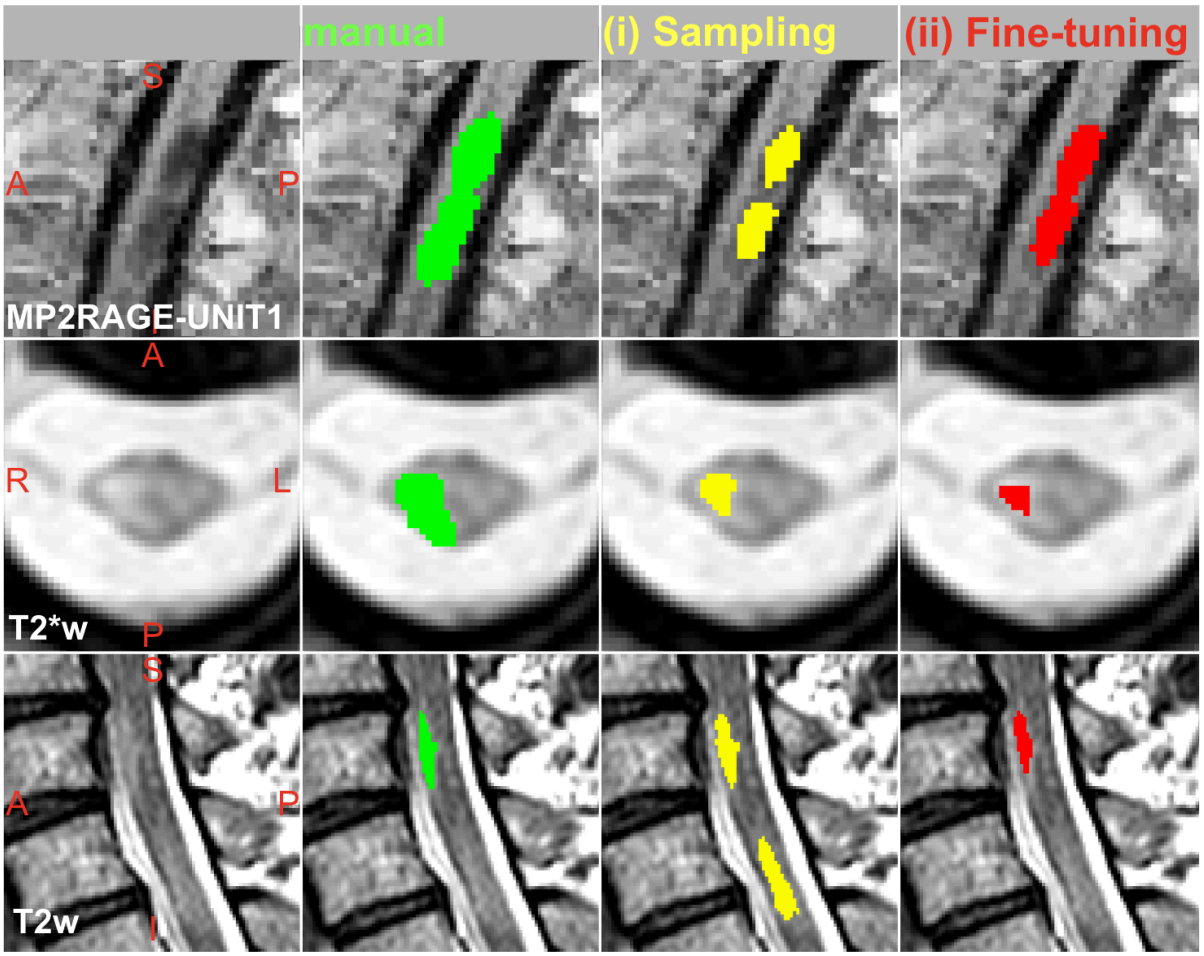

Figure 4: Qualitative examples of lesion segmentation for both the (i) sampling method and the (ii) fine-tuning method. Despite improved segmentation metrics, the (ii) Fine-tuning method seems to display under-segmentation patterns, while the (i) Sampling method even captures lesions missed during manual segmentation.

Data and Code Availability Statement:

No data are available for this abstract.

Code is available at https://github.com/ivadomed/ms-lesion-agnostic

References:

1. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. Lancet Neurol. 2018;17: 162–173.

2. Tent M. Revised McDonald criteria allow earlier and more precise MS diagnosis. In: Hartung H-P, editor. Medicom Conference Report ECTRIMS 2024. Baarn, the Netherlands: Medicom Medical Publishers; 2024. doi:10.55788/74e71270

3. Kerbrat A, Gros C, Badji A, Bannier E, Galassi F, Combès B, et al. Multiple sclerosis lesions in motor tracts from brain to cervical cord: spatial distribution and correlation with disability. Brain. 2020;143: 2089–2105.

4. Jackson-Tarlton CS, Flanagan EP, Messina SA, Barakat B, Ahmad R, Kantarci OH, et al. Progressive motor impairment from “critical” demyelinating lesions of the cervicomedullary junction. Mult Scler. 2023;29: 74–80.

5. Keegan BM, Kaufmann TJ, Weinshenker BG, Kantarci OH, Schmalstieg WF, Paz Soldan MM, et al. Progressive motor impairment from a critically located lesion in highly restricted CNS-demyelinating disease. Mult Scler. 2018;24: 1445–1452.

6. Cohen-Adad J, Alonso-Ortiz E, Abramovic M, Arneitz C, Atcheson N, Barlow L, et al. Generic acquisition protocol for quantitative MRI of the spinal cord. Nat Protoc. 2021;16: 4611–4632.

7. Saslow L, Li DKB, Halper J, Banwell B, Barkhof F, Barlow L, et al. An international standardized Magnetic Resonance Imaging protocol for diagnosis and follow-up of patients with multiple sclerosis: Advocacy, dissemination, and implementation strategies. Int J MS Care. 2020;22: 226–232.

8. Sastre-Garriga J, Pareto D, Battaglini M, Rocca MA, Ciccarelli O, Enzinger C, et al. MAGNIMS consensus recommendations on the use of brain and spinal cord atrophy measures in clinical practice. Nat Rev Neurol. 2020;16: 171–182.

9. Wattjes MP, Ciccarelli O, Reich DS, Banwell B, de Stefano N, Enzinger C, et al. 2021 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. Lancet Neurol. 2021;20: 653–670.

10. Lodé B, Hussein BR, Meurée C, Walsh R, Gaubert M, Lassalle N, et al. Evaluation of a deep learning segmentation tool to help detect spinal cord lesions from combined T2 and STIR acquisitions in people with multiple sclerosis. Eur Radiol. 2025. doi:10.1007/s00330-025-11541-0

11. Walsh R, Meurée C, Kerbrat A, Masson A, Hussein BR, Gaubert M, et al. Expert variability and deep learning performance in spinal cord lesion segmentation for multiple sclerosis patients. 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2023. pp. 463–470.

12. Polattímur R, Dandil E, Yildirim MS, Uluçay S, Şenol U. FractalSpiNet: Fractal-based U-net for automatic segmentation of cervical spinal cord and MS lesions in MRI. IEEE Access. 2024;12: 110955–110976.

13. Gros C, De Leener B, Badji A, Maranzano J, Eden D, Dupont SM, et al. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with

convolutional neural networks. Neuroimage. 2019;184: 901–915.

14. Naga Karthik E, McGinnis J, Wurm R, Ruehling S, Graf R, Valosek J, et al. Automatic segmentation of spinal cord lesions in MS: A robust tool for axial T2-weighted MRI scans. Radiology and Imaging. medRxiv; 2025. Available: https://www.medrxiv.org/content/10.1101/2025.01.22.25320959v1

15. Kamraoui RA, Ta V-T, Tourdias T, Mansencal B, Manjon JV, Coup P. DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. Med Image Anal. 2022;76: 102312.

16. Benveniste PL, Valošek J, Chen M, Molinier N, Eunyoung Lee L, Prat A, Vavasour Z, Tam R, Traboulsee A, Kolind S, Oh J, Cohen-Adad J. Automatic Segmentation of Spinal Cord MS Lesions Across Multiple Sites, Contrasts and Vendors. 9th annual Americas Committee for Treatment and Research in Multiple Sclerosis (ACTRIMS) Forum 2024, West Palm Beach, Florida. 2024.

17. Ulrich C, Wald T, Isensee F, Maier-Hein KH. Large scale supervised pretraining for traumatic brain injury segmentation. arXiv [cs.CV]. 2025. Available: http://arxiv.org/abs/2504.06741

18. Ulrich C. MultiTalentV2 Challenge Edition. Zenodo; 2024. doi:10.5281/ZENODO.13753413

19. Medina NL. model_seg_ms_mp2rage: Model repository for MS lesion segmentation on MP2RAGE data from University of Basel. Github; Available: https://github.com/ivadomed/model_seg_ms_mp2rage

20. Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024;21: 195–212.

21. Benveniste P-L, Lee LE, Prat A, Vavasour Z, Tam R, Traboulsee A, et al. Automatic multi-contrast MRI segmentation of spinal cord lesions. 10th annual Americas Committee for Treatment and Research in Multiple Sclerosis (ACTRIMS) Forum 2025, West Palm Beach, Florida. 2025.