Longitudinal tracking of multiple sclerosis lesions in the spinal cord: A validation study

Pierre–Louis Benveniste[1] PIERRE-LOUIS-2.BENVENISTE@POLYMTL.CA
Shannon Kolind[2] SHANNON.KOLIND@UBC.CA
Larry Lynd[2] LARRY.LYND@UBC.CA
Sarah A. Morrow[3] SARAH.MORROW@ALBERTAHEALTHSERVICES.CA
Jiwon Oh[4] JIWON.OH@UNITYHEALTH.TO
Alexandre Prat[5] A.PRAT@UMONTREAL.CA
Alice Schabas[2] ALICE.SCHABAS@UBC.CA
Penny Smyth[6] SMYTH@UALBERTA.CA
Roger Tam[2] ROGER.TAM@UBC.CA
Anthony Traboulsee[2] T.TRABOULSEE@UBC.CA
Hervé Lombaert[1] HERVE.LOMBAERT@POLYMTL.CA
Julien Cohen-Adad[1] JULIEN.COHEN-ADAD@POLYMTL.CA
[1] *Polytechnique Montréal, Montreal, QC, Canada* [2] *University of British Columbia, Vancouver, BC, Canada* [3] *University of Calgary, Calgary, AB, Canada* [4] *University of Toronto, Toronto, ON, Canada* [5] *Université de Montréal, Montreal, QC, Canada* [6] *University of Alberta, Edmonton, AB, Canada*

Editors: Under Review for MIDL 2026

Abstract

Longitudinal characterization of multiple sclerosis (MS) lesions remains constrained by the lack of frameworks capable of establishing consistent instance-level correspondences across time. Conventional segmentation approaches produce semantic lesion masks at each visit and therefore fail to capture the complex instance temporal patterns associated with lesion appearance, disappearance, splitting, or merging. This study presents a comparative evaluation of five strategies for automated tracking of spinal cord MS lesions in longitudinal MRI data from a multi-site cohort. The investigated strategies rely either on deformable registration or on a spinal anatomical reference system, and encompass overlap-based matching, coordinate-based Hungarian algorithm, gradient-boosted classification, and Siamese model classification. Tracking accuracy is quantified using instance-level true positives, false positives, and false negatives, allowing to assess the presence of one-to-many and many-to-one associations. Results show best performance for the registration-based overlap method. This study provides the first systematic analysis of lesion-instance correspondence in the spinal cord and outlines the strengths and limitations of registration-based and registration-free paradigms for longitudinal MS assessment. The code is available at: https://github.com/ivadomed/ms-lesion-agnostic

Keywords: Multiple Sclerosis, Longitudinal, MRI, Tracking, Lesion, Segmentation

1. Introduction

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system that produces demyelinating lesions in the brain and spinal cord (SC). Longitudinal monitoring of

these lesions with magnetic resonance imaging (MRI) is essential to detect change in disease course and adapt treatment appropriately, for example, in cases of conversion from relapsing-remitting to secondary progressive (Krieger et al., 2016). In routine practice, however, assessment of subtle lesion change remains approximate, i.e., there are limited measures of lesion growth and radiomic features. Instance lesion segmentation offers the opportunity to provide quantitative measures (e.g., lesion morphometrics, radiomic extraction), thereby improving clinical interpretation (Lipp et al., 2020).

Substantial progress has been made in segmentation of brain (Dereskewicz et al., 2025; Kaur et al., 2021; Wiltgen et al., 2024; Cerri et al., 2023) and SC MS lesions (Benveniste et al., 2025). Typically, these segmentations are encoded as a single-class value and are used to produce biomarkers such as total lesion load. However, instance-level metrics are needed to characterize specific lesions of interest, such as "critical" lesions (Ahmad et al., 2025). In addition, monitoring how each lesion evolves over time adds prognosis value. Achieving this not only requires accurate lesion delineation at each time point, but also precise tracking of lesions across longitudinal scans, that is, establishing a correspondence between lesion instances across multiple time points (Santoro-Fernandes et al., 2024). One difficulty is maintaining stable lesion identities over time despite substantial topological variability. Such correspondences are crucial in MS, where the evolution of the lesion frequently violates one-to-one temporal mapping assumptions. Accurate tracking must instead accommodate one-to-many and many-to-one correspondences arising from lesion splitting and merging events (Lassmann, 2018). Although several longitudinal lesion segmentation frameworks have been proposed for the brain (Carass et al., 2017; Kamraoui et al., 2022; Rokuss et al., 2025b; Krüger et al., 2020), they do not provide a consistent lesion-instance correspondence and thus cannot be used for tracking purposes.

To date, there is no dedicated methodology for the tracking of MS lesion in either the brain or the SC. In contrast, object tracking has been extensively explored in other domains, typically relying on one-to-one assignments derived from spatial overlap, trajectory regularity, or bipartite matching techniques such as the Hungarian algorithm (Bolme et al., 2010; Li et al., 2019, 2018; Rokuss et al., 2025a; Hering et al., 2021). Only a limited subset of medical imaging studies has considered asymmetric correspondences, addressing the more challenging one-to-many or many-to-one scenarios that arise when objects undergo fragmentation or fusion (Santoro-Fernandes et al., 2024; Di Veroli et al., 2023).

Tracking SC lesions longitudinally introduces specific methodological challenges. Unlike in the brain, where rigid or affine registration generally suffice for intra-participant alignment, SC imaging requires non-rigid registration due to the cord's articulated structure. Moreover, the SC is a thin structure ($\sim$1cm diameter), and MRI-visible lesions are small and susceptible to partial-volume effects. Even minimal misregistration can substantially alter lesion boundaries. This raises concerns regarding the reliability of registration-based approaches for temporal correspondence. Lesion tracking is further complicated by the pronounced inter- and intra-rater variability in SC lesion delineation, with reported fluctuations in lesion count reaching up to 50% (Walsh et al., 2023). Such variability reinforces the inadequacy of one-to-one correspondence assumptions and highlights the need for methods capable of handling uncertain or evolving lesion morphology.

Lesion evolution itself adds further complexity. Lesions may appear, resolve, split, or merge across time points. Figure 1A illustrates a case interpreted by the rater as a splitting

event, while Figure 1B shows a merging event. Diffusely abnormal white matter, extensively studied in both brain (Musall et al., 2024) and SC imaging (Bergers et al., 2002), may be segmented as a single confluent region or as multiple smaller clusters, depending on rater's interpretation. Figure 1C shows an example where a diffuse lesion was delineated differently across visits. These scenarios expose fundamental limitations of existing tracking techniques (Hering et al., 2021; Tang et al., 2022; Rokuss et al., 2025a; Yan et al.), which cannot deal with asymmetric correspondences.

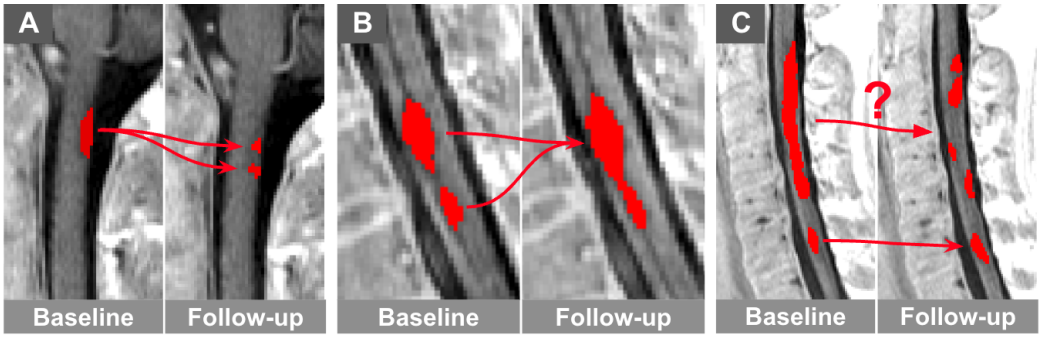

Figure 1: Examples of one–to–many, many–to–one, and uncertain longitudinal correspondences in SC MS lesions. A: A single focal lesion at baseline divides into two distinct lesions at follow-up, illustrating a one-to-many evolution. B: Two focal lesions at baseline merge, creating a confluent lesion at follow-up. C: A diffuse lesion is segmented as a single confluent region at baseline but is delineated as several smaller lesions at follow-up. That last example highlights how variability in the segmentation can produce markedly different longitudinal representations of the same underlying pathology.

In this work, we evaluate "standard" tracking methods leveraging registration and the Hungarian algorithm. We also investigate more complex methods using anatomical-based coordinate systems and deep learning methods to handle the complex, asymmetric temporal correspondence of evolving lesions. The objective of this work is to validate multiple automated frameworks for the complex tracking of SC MS lesion instances across longitudinal MRI using multi-site (n=5) clinical datasets.

2. Existing literature

Research on longitudinal lesion analysis in MS has predominantly focused on the brain, notably pushed in 2017 by the MS longitudinal lesion segmentation challenge (Carass et al., 2017). In this challenge, all scans were registered to a baseline scan, and human raters showed poor longitudinal consistency when blinded to time-order of scan. Top-performing methods relied on convolutional neural networks or classical machine-learning models. Following this initiative, several deep learning frameworks were proposed. Kamraoui et al. (Kamraoui et al., 2022) introduced a Siamese encoder coupled with a decoder trained to segment lesions from two affine-registered time points. Rokuss et al (Rokuss et al., 2024)

used dual encoders and feature-difference weighting blocks to emphasize temporal changes. An additional encoder–decoder formulation operating on pre-registered scans was described in (Krüger et al., 2020). All these approaches rely critically on image registration, which, according to Diaz-Hurtado et al. (Diaz-Hurtado et al., 2022), remains a mandatory component in all brain-based longitudinal pipelines. Importantly, these methods focus exclusively on longitudinal segmentation and do not perform lesion tracking. To date, no work has addressed longitudinal segmentation or tracking of MS lesions in the SC.

Beyond segmentation, longitudinal instance tracking has been investigated in several medical and non-medical settings, although not in MS. In computer vision, object tracking has been extensively studied, with high-performing methods leveraging Siamese networks, correlation filters, and others (Bolme et al., 2010; Li et al., 2018, 2019). These models are typically optimized for 2D, single-object trajectories, stable appearance characteristics, and minimal topological change, conditions that are incompatible with the evolution of MS lesions. In medical imaging, lesion tracking has mainly focused on tumor evolution. LesionLocator (Rokuss et al., 2025a) propagates a bounding box through time to track a tumor, implicitly assuming that the target remains a single connected entity. The DeepLesion dataset (Yan et al.) has supported the development of various approaches, including transformer-based architectures for temporal propagation (Tang et al., 2022) and registration-based matching strategies (Hering et al., 2021). These frameworks generally assume one-to-one correspondences across time points and are therefore unable to capture more complex patterns such as lesion splitting or merging.

Only a few studies have explicitly addressed asymmetric correspondences. Santoro-Fernandes et al. (Santoro-Fernandes et al., 2024) proposed a cancer lesion tracking method. To deal with lesion merging, they perform lesion clustering on dilated lesion masks using geometrical considerations such as lesion distance and longest lesion chord. Lesion groups are matched pairwise across scans using the Hungarian algorithm. Di Veroli et al. (Di Veroli et al., 2023) also investigated cancer lesion tracking. Lesions are matched using deformable registration and overlap. Lesion segmentations were dilated to compensate for small registration errors. Their overlap threshold was set at 10% and tracking performance was evaluated using true positives, false positives, and false negatives.

Overall, existing longitudinal MS segmentation frameworks do not perform lesion tracking. Only a handful of tracking methods in other domains can handle one-to-many and many-to-one correspondences. The specific challenges posed by SC MS lesions (small size, high morphological variability, rater variability, and dependence on non-rigid registration) underscore the need to develop tracking frameworks that can explicitly model complex lesion evolution, potentially leveraging deep learning to improve temporal consistency.

3. Methods

3.1. Dataset

The study used longitudinal data from the CanProCo cohort (Oh et al., 2021), each scanned at baseline (M0) and 12-month follow-up (M12). All participants with at least one lesion in the SC, for whom manual longitudinal lesion tracking had been completed, were included in this study (n=34, 13 males, age range at M0: 18-59 y.o.). Data were acquired across 5 hospitals (Calgary, Edmonton, Montreal, Toronto, and Vancouver) on Philips, Siemens,

or GE systems, using sagittal PSIR (n=29) or STIR (n=5) sequence, at 0.7×0.7×3 mm resolution.

3.2. Lesion segmentation

Lesion segmentations were generated using *sct_deepseg lesion_ms* (Benveniste et al., 2025) available in SCT (De Leener et al., 2017) v7.2. The model is a 5-fold U-Net optimized using the nnUNet framework. Inference was performed using a 5-fold ensemble with test-time augmentation based on mirroring along all axes.

Instance-level masks were extracted by connected-component analysis with 26-voxel connectivity.

3.3. Data split

The dataset was split on a participant basis into training (n=17), validation (n=8), and testing (n=9) sets. For each participant, all lesion pairs across M0 and M12 were enumerated. Each pair was labeled as positive if it represented the same longitudinal lesion instance, and negative otherwise. The resulting dataset was composed of 407 pairs for training (n=61 positive pairs), 172 pairs for validation (n=31 positive pairs), and 134 pairs for testing (n=24 positive pairs).

3.4. Lesion tracking

To explore longitudinal lesion correspondence, five tracking strategies were evaluated. Strategies 1-3 used a newly-introduced anatomical coordinate system, while strategies 4-5 relied on co-registration.

3.4.1. Strategy #1: Hungarian algorithm in the SC coordinate system

This approach defined each lesion solely by its centroid in (z, r, θ) coordinates, defined in the SC coordinate system.

In the SC coordinate system (as detailed in Figure 2), lesion centroids were expressed in cylindrical anatomical coordinates : z: continuous vertebral position along the labeled centerline obtained from *sct_get_centerline* and disc levels from TotalSpineSeg (Warszawer et al., 2024); r: radial distance from the centerline; θ: angular orientation relative to the anterior–posterior axis. These coordinates enabled registration-free comparison of lesion locations across time.

Longitudinal correspondences between M0 and M12 lesions were obtained using the Hungarian algorithm (Kuhn, 1955) applied to Euclidean distances between centroids. In brief, the Hungarian algorithm computes the optimal one-to-one assignment between two sets by minimizing the total assignment cost. To reflect the higher spatial reliability of the z axis in sagittal acquisitions, distances were anisotropically weighted with a 25:1 ratio between the z axis and the transverse plane. This empirically derived ratio was estimated from a randomly selected participant in the training set.

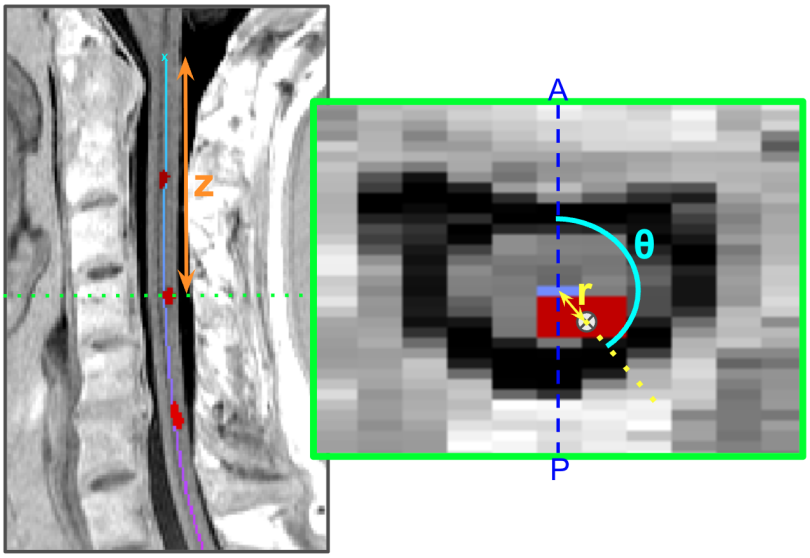

Figure 2: Illustration of cylindrical coordinates defining lesion centroids in the SC anatomical system. Left: sagittal PSIR image with lesions (in red) and the labeled centerline (in blue to pink). Right: axial slice corresponding to the green dotted line, where ⊗ denotes the centroid of the lesion. A: Anterior, P: Posterior

3.4.2. Strategy #2: XGBoost using lesion features in the SC coordinate system

This strategy formulated lesion tracking as a supervised binary classification task. An XGBoost model (Chen and Guestrin, 2016) was trained to discriminate matching and non-matching pairs of lesions from both time points of the same participants, using geometric descriptors. Each lesion was described by its SC-coordinate centroid (z, r, and θ), volume, and maximum Right-Left (RL)/Anterior-Posterior (AP)/Inferior-Superior (IS) diameters. Pair-level descriptors included component-wise displacements, Euclidean distance, and volume difference.

Model training used a log-loss objective. Hyperparameter optimization was performed using Bayesian optimization over 50 iterations with three-fold cross-validation and the average-precision score as the selection criterion. The final configuration retained after optimization was: colsample_bytree = 0.7998, learning_rate = 0.0555, max_depth = 7, min_child_weight = 7, scale_pos_weight = 9, and subsample = 0.8585.

Lesion pairs were assigned to the same tracks when their predicted probability exceeded 0.5. This strategy circumvented the one-to-one constraint inherent to Hungarian matching and avoided explicit registration steps.

3.4.3. Strategy #3: Siamese network using lesion features in the SC coordinate system

A Siamese model (Dey et al., 2017) was trained to discriminate lesion pairs of the same track. Each lesion was described by its SC-coordinate centroid (z, r, and θ), volume, and maximum RL/AP/IS diameters. Features were standardized prior to training using z-score normalization.

Both encoder branches consisted of a dense encoder ($128 \rightarrow 64 \rightarrow 32$) with ReLU activation and batch normalization. The resulting embeddings were compared using an L2 distance, followed by a sigmoid unit yielding a match probability. Training used a contrastive loss with margin 1, optimized via RMSprop (learning rate of 0.001). Because matching pairs were rare, batch-level class weights were used (0.59 for negatives, 3.28 for positives). Training proceeded for 500 epochs with a batch size of 20.

3.4.4. Strategy #4: Hungarian matching in the baseline scan reference space

Follow-up segmentations were first warped to the baseline image (M0) using *sct_register_multimodal* (parameters: $step = 0, type = label, dof = Tx\_Ty\_Tz; step = 1, type = im, algo = dl$) from SCT (De Leener et al., 2017). The registration method is based on a first step of pre-alignment of scans using disc levels and a second step performing registration using a deep learning model (Beal and Cohen-Adad, 2023). Intervertebral disc labeling, required for initialization, was performed using TotalSpineSeg (Warszawer et al., 2024). All registrations underwent visual quality control. Lesion centroids were computed in warped segmentation masks.

Lesion centroids were then computed in baseline space, and correspondences were determined through Hungarian matching based on centroid distances.

3.4.5. Strategy #5: Lesion overlapping in the baseline scan reference space

Follow-up segmentations were warped to the baseline scan space using the same registration method as in strategy #4. This strategy paired lesions across time points when their intersection-over-union (IoU) was non-zero. This voxel-overlap rule provided a direct, registration-based criterion for lesion continuity.

3.5. Evaluation metrics

Tracking performance was quantified by comparing the predicted lesion mappings with the ground-truth (GT) longitudinal mappings (Figure 3).

For each lesion at baseline, we computed the following evaluation metrics:

- True Positives (TP): the number of correctly identified follow-up lesions belonging to the same GT longitudinal track. Because a single baseline lesion may evolve into multiple follow-up lesions, TP may exceed 1 for a given baseline lesion.
- False Positives (FP): the number of follow-up lesions incorrectly associated with a baseline lesion.

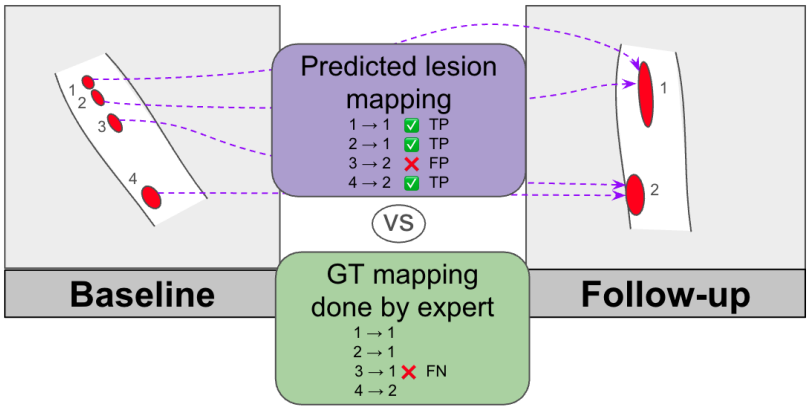

Figure 3: Evaluation of predicted lesion mappings. Predicted lesion mappings (purple box) are compared to GT lesion mappings done by an experts (green box). In this example, three lesions were correcly matched (TP=3), one lesion mapping was incorrect (FP=1) and one lesion mapping was missing (FN=1).

- False Negatives (FN): the number of missing follow-up assignments. These occurred either when a baseline lesion was not predicted or when the predicted mapping failed to link it to its corresponding follow-up lesion(s).

These metrics follow the same principles adopted in previous work on MS lesion tracking (Di Veroli et al., 2023; Santoro-Fernandes et al., 2024) and provide an interpretable basis for quantifying correct, spurious, and missed longitudinal correspondences.

4. Results

Figure 4 compares the five tracking strategies in a challenging case with a high lesion burden. In this example, Strategies #2 and #5 yielded accurate longitudinal correspondences, whereas Strategy #3 exhibited substantial mismatches. Strategies #1 and #4 were inherently constrained by the one-to-one assignment imposed by the Hungarian algorithm, limiting their ability to recover the many-to-one trajectory in this example.

Longitudinal tracking performance for each strategy was quantified using predicted lesion segmentations (Table 1). Overall, Strategy #5, combining registration to baseline with IoU-based correspondence, achieved the highest performance, reaching an F1 score of 0.98 on the test set. Results on all the data splits are detailed in Table 2 in Appendix A.

5. Discussion

In this study we assessed five strategies for the longitudinal tracking of MS lesions in the SC. The comparison of tracking approaches underscores the intrinsic difficulty of defining consistent lesion correspondences in the SC, where anatomical deformability, small and

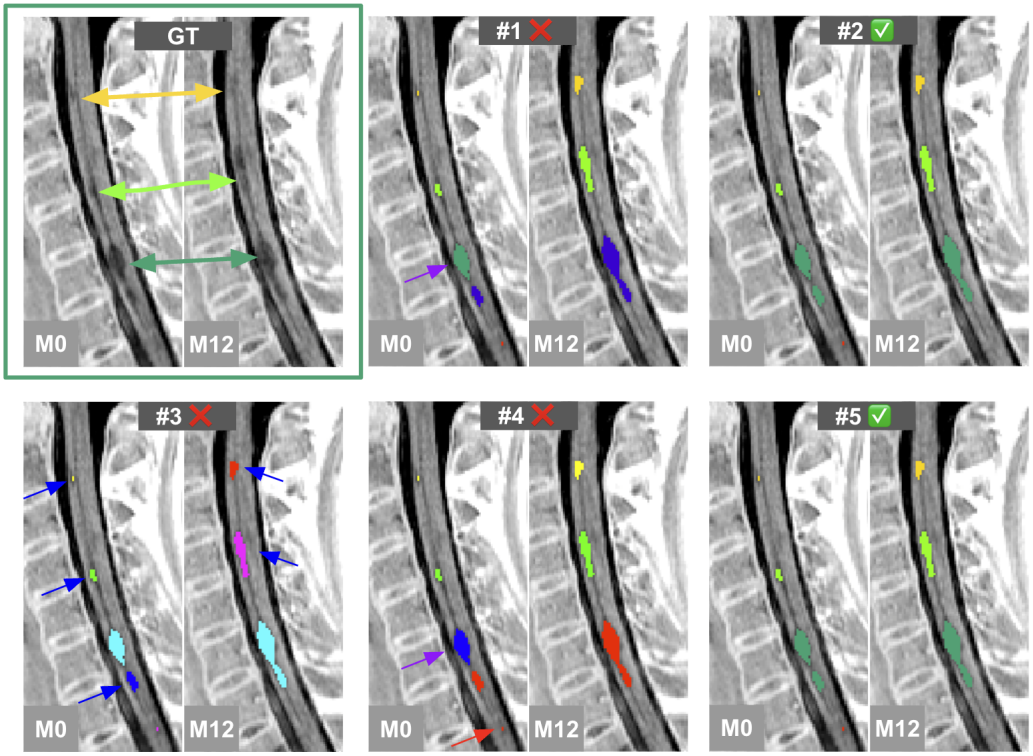

Figure 4: Qualitative comparison of the five tracking strategies on a PSIR scan from the Toronto site. Each image shows labeled lesion instances at M0 and M12, where lesions sharing the same color represent a predicted longitudinal trajectory. GT represents the correct lesion association between the two time points. →: false-positive mapping; →: incorrect correspondence; →: false-negative mapping.

Table 1: Longitudinal lesion-tracking performance across the five strategies on the test set.

	#1	#2	#3	#4	**#5**
TP ↑	21	21	11	21	**23**
FP ↓	5	1	4	5	**0**
FN ↓	3	3	13	3	**1**
Precision↑	0.81	0.95	0.73	0.81	**1.00**
Recall ↑	0.88	0.88	0.46	0.88	**0.96**
F1 score ↑	0.84	0.91	0.56	0.84	**0.98**

irregularly shaped lesions, substantial inter-rater variability, and frequent lesion merging or division complicate temporal association.

Among the evaluated strategies, the registration-based IoU method (#5) achieved the highest tracking performance. Allowing non-zero overlap proved advantageous for handling large volumetric changes, diffuse confluent lesions, or small registration errors. Nonetheless, IoU remains an imperfect criterion: it is sensitive to segmentation noise, deformations introduced by registration, and substantial longitudinal lesion evolution. As illustrated in Figure 4, subtle differences in segmentation can propagate into substantial discrepancies in downstream correspondence assignments.

Strategies relying on bipartite assignment via the Hungarian algorithm (#1 and #4) were structurally constrained, as they enforce one-to-one mappings. This assumption is incompatible with the one-to-many and many-to-one trajectories commonly observed in MS. Their lower performance reflects this inherent limitation. Despite this constraint, the coordinate-based Hungarian approach (#1) performed surprisingly well. This suggests that the proposed anatomical coordinate system captures meaningful spatial structure that can support tracking.

Feature-based approaches (XGBoost and Siamese network) using the SC coordinate system produced mixed results. The XGBoost model (#2) achieved competitive performance close to the IoU-based strategy (#5), demonstrating that geometric descriptors alone can encode salient longitudinal relationships without requiring deformable registration. Its tendency toward over-confident predictions, reflected in higher false positives than Strategy #5, may indicate that the model captures patterns associated with early-stage lesion merging. The Siamese architecture (#3), despite extensive experimentation—including alternative normalization schemes, deeper or wider encoders, different distance metrics, multiple loss formulations, and a range of optimizers—did not yield gains beyond simpler models. This is likely attributable to the limited dataset size and the restricted discriminative power of the available geometric descriptors, which are not fit for this type of architecture.

This study relied on a disc-based SC reference system, in which vertebral levels provide an indirect surrogate for local SC anatomy. This approach remains sensitive to head and neck positioning differences across sessions: changes in cervical curvature can shift the cord relative to the vertebrae, thereby biasing coordinate estimates. Recent developments using the pontomedullary junction as a reference (Bédard et al., 2023) offer a promising alternative by anchoring coordinates directly to intrinsic SC anatomy.

Although the IoU-based method (#5) performed best, its reliability is tied to the quality of the underlying registration fields and the accuracy of disc labeling, the latter being produced by the TotalSpineSeg algorithm (Warszawer et al., 2024). Failures in disc detection or registration directly impact strategies operating either in the registered space or in the SC coordinate system, as both depend on accurate initialization.

A persistent challenge highlighted in this work is the inadequacy of existing instance-level evaluation metrics for assessing longitudinal lesion tracking. Metrics commonly used in instance segmentation, such as Difference in Count, Mean Precision, Mean Coverage Loss, or Average Best Overlap (Molina et al., 2025), do not capture temporal correspondence. Similarly, multi-object tracking metrics (MOTA, MOTP) (Bernardin and Stiefelhagen, 2008) are inappropriate because they either quantify segmentation quality or assume strictly one-to-one temporal mappings. Even the Panoptic Quality metric used in recent MS lesion

instance segmentation work (Wynen et al., 2024) does not extend naturally to longitudinal association. This underscores the need for dedicated metrics tailored to tracking in cases of complex correspondences.

Alternative definitions of lesion instances more complex than connected components, such as Confluent Lesion Splitting (CLS), have been proposed to better represent confluent lesions (Wynen et al., 2024; Lou et al., 2021). While such approaches might conceptually align with the challenges of longitudinal tracking, prior evidence (Wynen et al., 2024) indicates that CLS can degrade segmentation quality relative to connected components, suggesting a trade-off between anatomical realism and segmentation accuracy.

A deliberate methodological choice in this study was to restrict tracking criteria to spatial and geometric descriptors in the SC coordinate system, without exploiting lesion shape or appearance in the MRI scan. These features, although informative, risk introducing systematic errors in cases of substantial morphological evolution, MRI scanner change, or acquisition parameter changes. Restricting the feature set also ensures independence from MRI contrast and acquisition site, promoting the generalizability of the method.

The evaluation protocol assessed tracking accuracy at the level of pairwise correspondences between time-adjacent lesions rather than full lesion trajectories, thereby simplifying comparison across strategies. This link-based evaluation is consistent with prior work in lesion tracking, such as (Santoro-Fernandes et al., 2024) in oncological imaging.

6. Conclusion

This study provided the first investigation of longitudinal instance-level tracking of MS lesions in the SC. By formulating lesion tracking as a correspondence problem rather than a segmentation task, the work addressed a gap unfilled by existing longitudinal MS frameworks. In particular, we address the difficult problem of one-to-many and many-to-one associations in object tracking. The results demonstrated that registration-based overlap yielded the highest overall performance, indicating that spatial alignment remains a powerful—although imperfect—mechanism for temporal lesion association. Yet, the competitive performance of the SC–coordinate–based XGBoost model showed that meaningful longitudinal information can be extracted without deformable registration, suggesting an opportunity for registration-free approaches. Future research should validate these findings in larger cohorts and explore the impact of tracking lesions individually in a clinical setting.

Acknowledgments

Funded by the Canada Research Chair in Quantitative Magnetic Resonance Imaging [CRC-2020-00179], the Canadian Institute of Health Research [PJT-190258, PJT-203803], the Canada Foundation for Innovation [32454, 34824], the Fonds de Recherche du Québec - Santé [322736, 324636, 370582], the UNIQUE Excellence Scholarship, the Natural Sciences and Engineering Research Council of Canada [RGPIN-2019-07244], the Canada First Research Excellence Fund (IVADO and TransMedTech), the Courtois NeuroMod project, the Quebec BioImaging Network [5886, 35450], INSPIRED (Spinal Research, UK; Wings for Life, Austria; Craig H. Neilsen Foundation, USA), Mila - Tech Transfer Funding Program.

References

Rowaid Ahmad, Caitlin S Jackson-Tarlton, Eoin P Flanagan, Steven Messina, W Scott Harmsen, Tuba Chaudhry, Nabeela Nathoo, W Oliver Tobin, Michel Toledano, Sean J Pittock, Claudia F Lucchinetti, Brian G Weinshenker, Orhun H Kantarci, and B Mark Keegan. Critical demyelinating lesions in progressive multiple sclerosis: a prospective observational study. *J. Neurol.*, 272(10):677, October 2025.

E Beal and J Cohen-Adad. Contrast-agnostic deep learning–based registration pipeline: Validation in spinal cord multimodal MRI data. *Aperture Neuro*, 3(2023):1–21, July 2023.

Pierre-Louis Benveniste, Lisa Eunyoung Lee, Alexandre Prat, Zachary Vavasour, Roger Tam, Anthony Traboulsee, Shannon Kolind, Jiwon Oh, Michelle Chen, Charidimos Tsagkas, Christina Granziera, Nilser Laines Medina, Mark Muhlau, Jan Kirschke, Julian McGinnis, Daniel S Reich, Christopher Hemond, Virginie Callot, Sarah Demortière, Bertrand Audoin, Govind Nair, Massimo Filippi, Paola Valsasina, Maria A Rocca, Olga Ciccarelli, Marios Yiannakas, Tobias Granberg, Russell Ouellette, Shahamat Tauhid, Rohit Bakshi, Caterina Mainero, Constantina Andrada Treaba, Anne Kerbrat, Elise Bannier, Gilles Edan, Pierre Labauge, Kristin P O'Grady, Seth A Smith, Timothy M Shepherd, Erik Charlson, Jean-Christophe Brisset, Jason Talbott, Yaou Liu, Hervé Lombaert, and Julien Cohen-Adad. Reinforcing the generalizability of spinal cord multiple sclerosis lesion segmentation models. In *Proceedings of the 41st Annual Scientific Meeting of the ESMRMB*, 2025.

Elisabeth Bergers, Joost C J Bot, Paul van der Valk, Jonas A Castelijns, Geert J Lycklama a Nijeholt, Wouter Kamphorst, Chris H Polman, Erwin L A Blezer, Klaas Nicolay, Rivka Ravid, and Frederik Barkhof. Diffuse signal abnormalities in the spinal cord in multiple sclerosis: direct postmortem in situ magnetic resonance imaging correlated with in vitro high-resolution magnetic resonance imaging and histopathology. *Ann. Neurol.*, 51(5): 652–656, May 2002.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008: 1–10, 2008.

Dav Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2544–2550. IEEE, June 2010.

Sandrine Bédard, Maxime Bouthillier, and Julien Cohen-Adad. Pontomedullary junction as a reference for spinal cord cross-sectional area: validation across neck positions. *Sci. Rep.*, 13(1):13527, August 2023.

Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, Manuel Jorge Cardoso, Niamh Cawley, Olga Ciccarelli, Claudia A M Wheeler-Kingshott, Sébastien Ourselin, Laurence Catanese, Hrishikesh Deshpande, Pierre Maurel, Olivier Commowick, Christian Barillot, Xavier Tomas-Fernandez, Simon K Warfield, Suthirth Vaidya, Abhijith Chunduru, Ramanathan Muthuganapathy, Ganapathy Krishnamurthi, Andrew Jesson, Tal Arbel, Oskar Maier, Heinz Handels, Leonardo O Iheme, Devrim Unay, Saurabh Jain, Diana M Sima, Dirk Smeets, Mohsen Ghafoorian, Bram Platel, Ariel Birenbaum, Hayit Greenspan, Pierre-Louis Bazin, Peter A Calabresi, Ciprian M Crainiceanu, Lotta M Ellingsen, Daniel S Reich, Jerry L Prince, and Dzung L Pham. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *Neuroimage*, 148:77–102, March 2017.

Stefano Cerri, Douglas N Greve, Andrew Hoopes, Henrik Lundell, Hartwig R Siebner, Mark Mühlau, and Koen Van Leemput. An open-source tool for longitudinal whole-brain and white matter lesion segmentation. *NeuroImage Clin.*, 38(103354):103354, March 2023.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, August 2016. ACM.

Benjamin De Leener, Simon Lévy, Sara M Dupont, Vladimir S Fonov, Nikola Stikov, D Louis Collins, Virginie Callot, and Julien Cohen-Adad. SCT: Spinal cord toolbox, an open-source software for processing spinal cord MRI data. *Neuroimage*, January 2017.

Emma Dereskewicz, Francesco La Rosa, Jonadab Dos Santos Silva, Edward Sizer, Amit Kohli, Maxence Wynen, William A Mullins, Pietro Maggi, Sarah Levy, Kamso Onyemeh, Batuhan Ayci, Andrew J Solomon, Jakob Assländer, Omar Al-Louzi, Daniel S Reich, James Sumowski, and Erin S Beck. FLAMeS: A robust deep learning model for automated multiple sclerosis lesion segmentation. *medRxiv*, May 2025.

Sounak Dey, Anjan Dutta, J Ignacio Toledo, Suman K Ghosh, Josep Llados, and Umapada Pal. SigNet: Convolutional siamese network for writer independent offline signature verification. *arXiv [cs.CV]*, July 2017.

Beniamin Di Veroli, Richard Lederman, Jacob Sosna, and Leo Joskowicz. Graph-theoretic automatic lesion tracking and detection of patterns of lesion changes in longitudinal CT studies. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 106–115. Springer Nature Switzerland, Cham, 2023.

Marcos Diaz-Hurtado, Eloy Martínez-Heras, Elisabeth Solana, Jordi Casas-Roma, Sara Llufriu, Baris Kanber, and Ferran Prados. Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review. *Neuroradiology*, 64(11):2103–2117, November 2022.

Alessa Hering, F Peisen, T Amaral, S Gatidis, T Eigentler, A Othman, and J Moltz. Whole-body soft-tissue lesion tracking and segmentation in longitudinal CT imaging studies. *MIDL*, 143:312–326, January 2021.

Reda Abdellah Kamraoui, Boris Mansencal, José V Manjon, and Pierrick Coupé. Longitudinal detection of new MS lesions using deep learning. *Front. Neuroimaging*, 1:948235, August 2022.

Amrita Kaur, Lakhwinder Kaur, and Ashima Singh. State-of-the-art segmentation techniques and future directions for multiple sclerosis brain lesions. *Arch. Comput. Methods Eng.*, 28(3):951–977, May 2021.

Stephen C Krieger, Karin Cook, Scott De Nino, and Madhuri Fletcher. The topographical model of multiple sclerosis: A dynamic visualization of disease course. *Neurol. Neuroimmunol. Neuroinflamm.*, 3(5):e279, October 2016.

Julia Krüger, Roland Opfer, Nils Gessert, Ann-Christin Ostwaldt, Praveena Manogaran, Hagen H Kitzler, Alexander Schlaefer, and Sven Schippling. Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. *NeuroImage Clin.*, 28(102445):102445, September 2020.

H W Kuhn. The hungarian method for the assignment problem. *Nav. Res. Logist. Q.*, 2 (1-2):83–97, March 1955.

Hans Lassmann. Multiple sclerosis pathology. *Cold Spring Harb. Perspect. Med.*, 8(3), March 2018.

Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018.

Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019.

Ilona Lipp, Greg D Parker, Emma C Tallantyre, Alex Goodall, Steluta Grama, Eleonora Patitucci, Phoebe Heveron, Valentina Tomassini, and Derek K Jones. Tractography in the presence of multiple sclerosis lesions. *Neuroimage*, 209(116471):116471, April 2020.

Carolyn Lou, Pascal Sati, Martina Absinta, Kelly Clark, Jordan D Dworkin, Alessandra M Valcarcel, Matthew K Schindler, Daniel S Reich, Elizabeth M Sweeney, and Russell T Shinohara. Fully automated detection of paramagnetic rims in multiple sclerosis lesions on 3T susceptibility-based MR imaging. *NeuroImage Clin.*, 32(102796):102796, August 2021.

José M Molina, Juan P Llerena, Luis Usero, and Miguel A Patricio. Advances in instance segmentation: Technologies, metrics and applications in computer vision. *Neurocomputing*, 625(129584):129584, April 2025.

Benjamin C Musall, Refaat E Gabr, Yanyu Yang, Arash Kamali, John A Lincoln, Michael A Jacobs, Vi Ly, Xi Luo, Jerry S Wolinsky, Ponnada A Narayana, and Khader M Hasan. Detection of diffusely abnormal white matter in multiple sclerosis on multiparametric brain MRI using semi-supervised deep learning. *Sci. Rep.*, 14(1):17157, July 2024.

Jiwon Oh, Nathalie Arbour, Fabrizio Giuliani, Melanie Guenette, Shannon Kolind, Larry Lynd, Ruth Ann Marrie, Luanne M Metz, Scott B Patten, Alexandre Prat, Alice Schabas, Penelope Smyth, Roger Tam, Anthony Traboulsee, and V Wee Yong. The canadian prospective cohort study to understand progression in multiple sclerosis (CanProCo): rationale, aims, and study design. *BMC Neurol.*, 21(1):418, October 2021.

Maximilian Rokuss, Yannick Kirchhoff, Saikat Roy, Balint Kovacs, Constantin Ulrich, Tassilo Wald, Maximilian Zenk, Stefan Denner, Fabian Isensee, Philipp Vollmuth, Jens Kleesiek, and Klaus Maier-Hein. Longitudinal segmentation of MS lesions via temporal difference weighting. *arXiv [eess.IV]*, September 2024.

Maximilian Rokuss, Yannick Kirchhoff, Seval Akbal, Balint Kovacs, Saikat Roy, Constantin Ulrich, Tassilo Wald, Lukas T Rotkopf, Heinz-Peter Schlemmer, and Klaus Maier-Hein. LesionLocator: Zero-shot universal tumor segmentation and tracking in 3D whole-body imaging. *arXiv [cs.CV]*, pages 30872–30885, February 2025a.

Maximilian R Rokuss, Yannick Kirchhoff, Saikat Roy, Balint Kovacs, Constantin Ulrich, Tassilo Wald, Maximilian Zenk, Stefan Denner, Fabian Isensee, Philipp Vollmuth, Jens Kleesiek, and Klaus Maier-Hein. Longitudinal segmentation of MS lesions via temporal difference weighting. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 64–74. Springer Nature Switzerland, Cham, 2025b.

Victor Santoro-Fernandes, Daniel T Huff, Luciano Rivetti, Alison Deatsch, Brayden Schott, Scott B Perlman, and Robert Jeraj. An automated methodology for whole-body, multimodality tracking of individual cancer lesions. *Phys. Med. Biol.*, 69(8):085012, April 2024.

Wen Tang, Han Kang, Haoyue Zhang, Pengxin Yu, Corey W Arnold, and Rongguo Zhang. Transformer lesion tracker. *arXiv [cs.CV]*, June 2022.

Ricky Walsh, Cédric Meurée, Anne Kerbrat, Arthur Masson, Burhan Rashid Hussein, Malo Gaubert, Francesca Galassi, and Benoit Combés. Expert variability and deep learning performance in spinal cord lesion segmentation for multiple sclerosis patients. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, June 2023.

Yehuda Warszawer, Nathan Molinier, Jan Valošek, Sandrine Bédard, Emanuel Shirbint, Pierre-Louis Benveniste, Anat Achiron, Arman Eshaghi, and Julien Cohen-Adad. TotalSpineSeg: Robust segmentation and labeling of vertebrae, intervertebral discs, spinal cord, and spinal canal in MRI images using nnU-net and iterative algorithm, 2024.

Tun Wiltgen, Julian McGinnis, Sarah Schlaeger, Florian Kofler, Cuici Voon, Achim Berthele, Daria Bischl, Lioba Grundl, Nikolaus Will, Marie Metz, David Schinz, Dominik Sepp, Philipp Prucker, Benita Schmitz-Koep, Claus Zimmer, Bjoern Menze, Daniel Rueckert, Bernhard Hemmer, Jan Kirschke, Mark Mühlau, and Benedikt Wiestler. LST-AI: A deep learning ensemble for accurate MS lesion segmentation. *NeuroImage Clin.*, 42(103611):103611, April 2024.

Maxence Wynen, P Gordaliza, Anna Stölting, Pietro Maggi, M Cuadra, and Benoit Macq. Lesion instance segmentation in multiple sclerosis: Assessing the efficacy of statistical lesion splitting. *ISMRM Annual Meeting*, 2024.

Ke Yan, Xiaosong Wang, and Le Lu. DeepLesion dataset.

Appendix A. Results on all the data splits

Table 2: Longitudinal lesion-tracking performance across the five strategies.

	#1		#2		#3		#4		#5	
	Train/Val	Test	Train/Val	Test	Train/Val	Test	Train/Val	Test	Train/Val	Test
TP ↑	53/23	21	–	21	–	11	53/22	21	55/29	23
FP ↓	13/8	5	–	1	–	4	7/9	5	0/0	0
FN ↓	9/7	3	–	3	–	13	8/8	3	6/1	1
Precision↑	0.80/0.74	0.81	–	0.95	–	0.73	0.88/0.71	0.81	1.00/1.00	1.00
Recall ↑	0.85/0.77	0.88	–	0.88	–	0.46	0.87/0.73	0.88	0.97/0.97	0.96
F1 score ↑	0.83/0.75	0.84	–	0.91	–	0.56	0.88/0.72	0.84	0.98/0.98	0.98