SCIseg: Automatic Segmentation of T2-weighted Hyperintense Lesions in Spinal Cord Injury

Naga Karthik Enamundram, MSc*†[1,2], Jan Valosek, PhD*[1,2,3,4], Andrew C. Smith, PT, DPT, PhD[5], Dario Pfyffer, PhD[6,7], Simon Schading-Sassenhausen, MSc[6], Lynn Farner, MSc[6], Kenneth A. Weber II, DC, PhD[7], Patrick Freund, MD, PhD[6,8], Julien Cohen-Adad, PhD[1,2,9,10]

*Shared co-first authorship - authors contributed equally

†Corresponding author (email: naga-karthik.enamundram@polymtl.ca)

1. NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada
2. Mila - Quebec AI Institute, Montreal, QC, Canada
3. Department of Neurosurgery, Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc, Czechia
4. Department of Neurology, Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc, Czechia
5. Department of Physical Medicine and Rehabilitation Physical Therapy Program, University of Colorado School of Medicine, Aurora, Colorado, USA
6. Spinal Cord Injury Center, Balgrist University Hospital, University of Zürich, Zürich, Switzerland
7. Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Stanford, California, USA
8. Department of Neurophysics, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
9. Functional Neuroimaging Unit, CRIUGM, Université de Montréal, Montreal, QC, Canada
10. Centre de Recherche du CHU Sainte-Justine, Université de Montréal, Montreal, QC, Canada

Summary

Automatic segmentation of the spinal cord and T2-weighted hyperintense lesions in spinal cord injury on MRI scans across different treatment strategies, lesion etiologies, sites, scanner manufacturers, and heterogeneous image resolutions.

Key Results

- An open-source, automatic method, `SCIseg`, was trained on a dataset of 191 spinal cord injury patients from three sites for the segmentation of spinal cord and T2-weighted hyperintense lesions.
- `SCIseg` generalizes across traumatic and non-traumatic lesions, scanner manufacturers, and heterogeneous image resolutions, enabling the automatic extraction of lesion morphometrics in large multi-site cohorts.
- Morphometrics derived from the automatic predictions showed no statistically significant difference when compared with manual ground truth, implying reliability in `SCIseg`'s predictions.

This work has been submitted to Radiology: Artificial Intelligence for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

Abstract

Background: Quantitative MRI biomarkers in spinal cord injury (SCI) can help understand the extent of the focal injury. However, due to the lack of automatic segmentation methods, these biomarkers are derived manually, which is a time-consuming process prone to intra- and inter-rater variability, thus limiting large multi-site studies and translation to clinical workflows.

Purpose: To develop a deep learning tool for the automatic segmentation of T2-weighted hyperintense lesions and the spinal cord in SCI patients.

Material and Methods: This retrospective study included a cohort of SCI patients from three sites enrolled between July 2002 and February 2023 who underwent clinical MRI examination. A deep learning model, `SCIseg`, was trained on T2-weighted images with heterogeneous image resolutions (isotropic, anisotropic), and orientations (axial, sagittal) acquired using scanners from different manufacturers (Siemens, Philips, GE) and different field strengths (1T, 1.5T, 3T) for the automatic segmentation of SCI lesions and the spinal cord. The proposed method was visually and quantitatively compared with other open-source baseline methods. Quantitative biomarkers (lesion volume, lesion length, and maximal axial damage ratio) computed from manual ground-truth lesion masks and automatic `SCIseg` predictions were correlated with clinical scores (pinprick, light touch, and lower extremity motor scores). A between-group comparison was performed using the Wilcoxon signed-rank test.

Results: MRI data from 191 SCI patients (mean age, 48.1 years ± 17.9 [SD]; 142 males) were used for training. Compared to existing methods, `SCIseg` achieved the best segmentation performance for both the cord and lesions and generalized well to both traumatic and non-traumatic SCI patients. `SCIseg` is open-source and accessible through the Spinal Cord Toolbox.

Conclusion: Automatic segmentation of intramedullary lesions commonly seen in traumatic SCI replaces the tedious manual annotation process and enables the extraction of relevant lesion morphometrics in large cohorts. The proposed model generalizes across lesion etiologies (traumatic, ischemic), scanner manufacturers and heterogeneous image resolutions.

Keywords

Spinal Cord, Trauma, Segmentation, MR-Imaging, Supervised learning, Convolution Neural Networks (CNN)

List of Abbreviations

DCM = degenerative cervical myelopathy
DL = deep learning
RVE = relative volume error
SC = spinal cord
SCI = spinal cord injury

1. Introduction

Traumatic spinal cord injury (SCI) results from acute damage to the spinal cord (SC) due to external physical factors (1). The majority of traumatic SCI patients sustain permanent neurological deficits such as motor and autonomic dysfunction with devastating physical and social consequences (1). Conventional MRI provides macrostructural information about the level of injury, intramedullary and extramedullary abnormalities (e.g., edema and hemorrhage), and allows the evaluation of soft tissue structures (1,2). Importantly, MRI-derived quantitative biomarkers, namely, intramedullary lesion length and lesion volume, have demonstrated associations with the neurological prognosis of traumatic SCI patients (3–8).

Despite recent advances in automatic SC MRI processing (9–12), robust methods for automatic quantitative MRI biomarker identification in traumatic SCI are still missing. As a result, most studies involve manual identification of these biomarkers (6,8,13–16), which is a time-consuming process susceptible to inter-rater variability, making it less reproducible in multi-site studies. Deep learning (DL) can improve the diagnosis and prognostication in SCI by automating the lesion annotation process, thereby reducing rater-specific biases and facilitating the analysis of large SCI cohorts across sites (17–19). Indeed, quantitative SCI lesion biomarkers derived from DL-based automatic segmentations have been shown to correlate well with clinical measures of motor impairment (20). However, despite its numerous advantages, DL has not been sufficiently explored in the context of traumatic SCI (18), with no open-source methods existing to date. This suggests a need for an automatic biomarker identification method that deals with the complex pathophysiology of traumatic SCI patients, generalizes to multiple sites and is easily accessible by researchers.

In this study, we introduced a DL-based method, *SCIseg*, for automatically segmenting the spinal cord and T2-weighted hyperintense lesions in volumetric SCI images. We performed correlation analyses between the automatically derived lesion biomarkers and clinical scores. To our best knowledge, this is the first open-source method for the 3D segmentation of both the SC and lesions, which can perform well across treatment strategies (pre-operative/post-operative), lesion etiologies (traumatic/ischemic), sites, scanner manufacturers, heterogeneous image resolutions, and fields of view. Importantly, `SCIseg` generalizes well to non-traumatic SCI (i.e., degenerative cervical myelopathy, DCM), obtaining reasonable segmentations for DCM lesions.

2. Materials and Methods

2.1 Study Design and Participants

This retrospective study included a cohort of SCI patients from three sites enrolled between July 2002 and February 2023. All patients provided written informed consent following Institutional Review Board approval and the Declaration of Helsinki. Demographic characteristics are provided in Table 1. Patients from site 2 were clinically assessed using the international standards for the neurological classification of SCI (ISNCSCI) protocol to obtain light touch, pinprick, and lower extremity motor scores. Eight patients from site 1 were followed up with additional MRI examinations. The inclusion criteria were: post-traumatic SCI, clinical MRI available for analyses, and completed enrollment. Exclusion criteria were: concurrent traumatic brain injury beyond concussion, and significant pre-existing neurological history (i.e., multiple sclerosis, cerebrovascular stroke). Patients from all sites were reported previously (6,8,13,21,22). These studies focused on understanding the consequences of SCI in the context of their predictive relationships with motor and sensory functions while using manually segmented lesion masks. In contrast, our study presents an automatic tool for segmenting T2-weighted hyperintensities, potentially aiding future clinical studies in SCI.

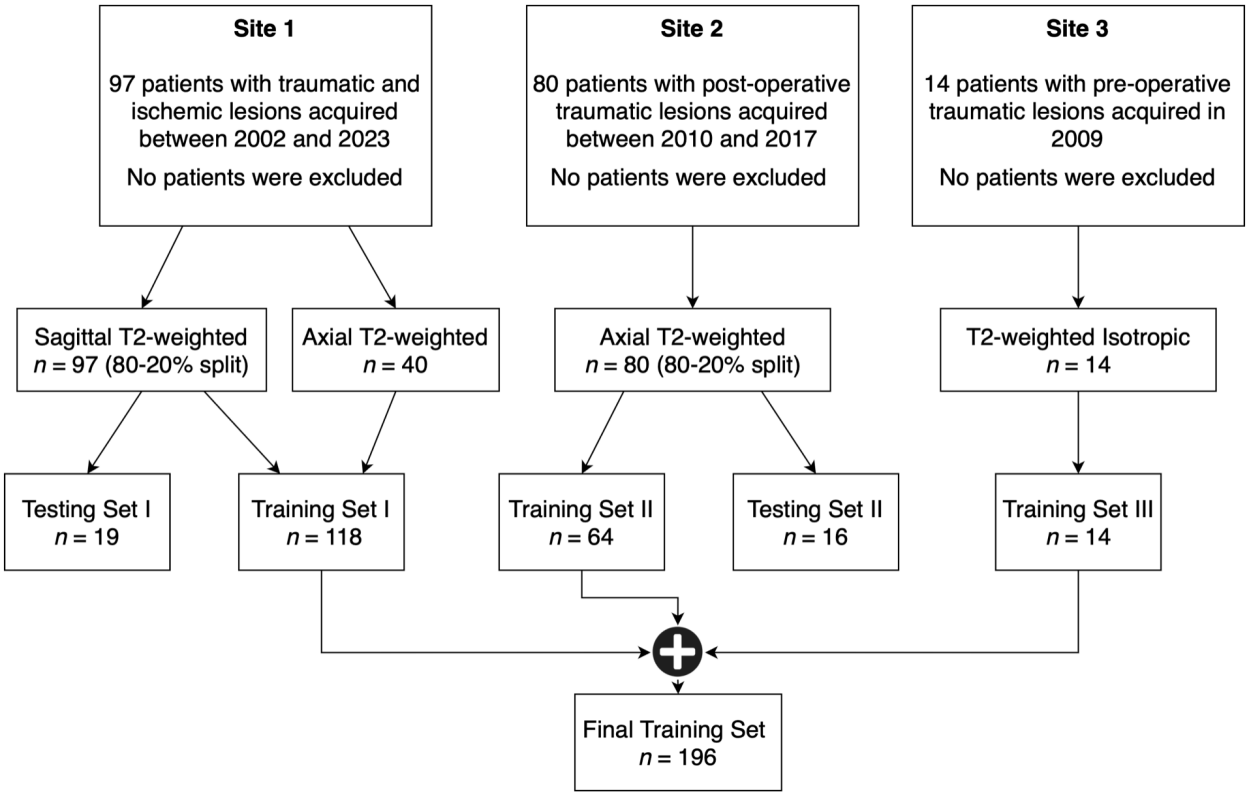

Figure 1: Study Flowchart. The data included patient cohorts from three sites with heterogeneous image resolutions, orientations, and lesion etiologies. The validation set is included within the final training set. Models were evaluated independently on the test sets of Site 1 and Site 2. Please refer to Table 1 for details on the MRI vendors and field strengths.

2.2 MRI Data

T2-weighted MRI scans with varying lesion types (traumatic/ischemic), voxel sizes, and orientations (sagittal/axial) were used (Table 1). Four raters segmented SC and lesions using JIM and FSLeyes image viewers. Due to the absence of ground-truth SC masks for site 1, we used the `sct_deepseg_sc` (23) to initially segment the SC, followed by manual corrections with the consensus of two raters (with 2 and 7 years of experience, respectively), if necessary.

2.3 Deep Learning Training Protocol

The model was trained in three phases (Figure 2). In the initial phase, a baseline segmentation model was trained using a labeled dataset of 78 subjects with T2-weighted sagittal images (site 1) and 64 subjects with T2-weighted axial images (site 2). We used the *region-based training* strategy of nnUNet (24), where the model initially segments the SC and then localizes itself on the SC to segment the T2-weighted hyperintense lesions subsequently. Default data augmentation methods by nnUNet were used. All images were preprocessed with RPI orientation and Z-score normalization. The model was trained for 1000 epochs, with a batch size of 2 using the stochastic gradient descent optimizer with a polynomial learning rate scheduler.

For the second phase, we used the human-in-the-loop active learning strategy (25) to gradually include axial T2-weighted images from site 1 in the training dataset. Using the phase 1 baseline model, we generated initial SC and lesion predictions for unlabeled axial images from site 1. A subset of predicted segmentations underwent quality control, with two raters manually correcting if needed. These refined segmentations were then added to the training dataset, resulting in the inclusion of 40 scans and leading to a total of 182 scans in the training set.

To further improve our model's generalization capabilities to a wide range of image resolutions, we added a new dataset from site 3 containing 14 isotropic T2-weighted sagittal scans of pre-operative traumatic SCI patients in the third training phase. In summary, the final dataset consisting of 196 scans from three sites was used to train the final model using the region-based training procedure described above.

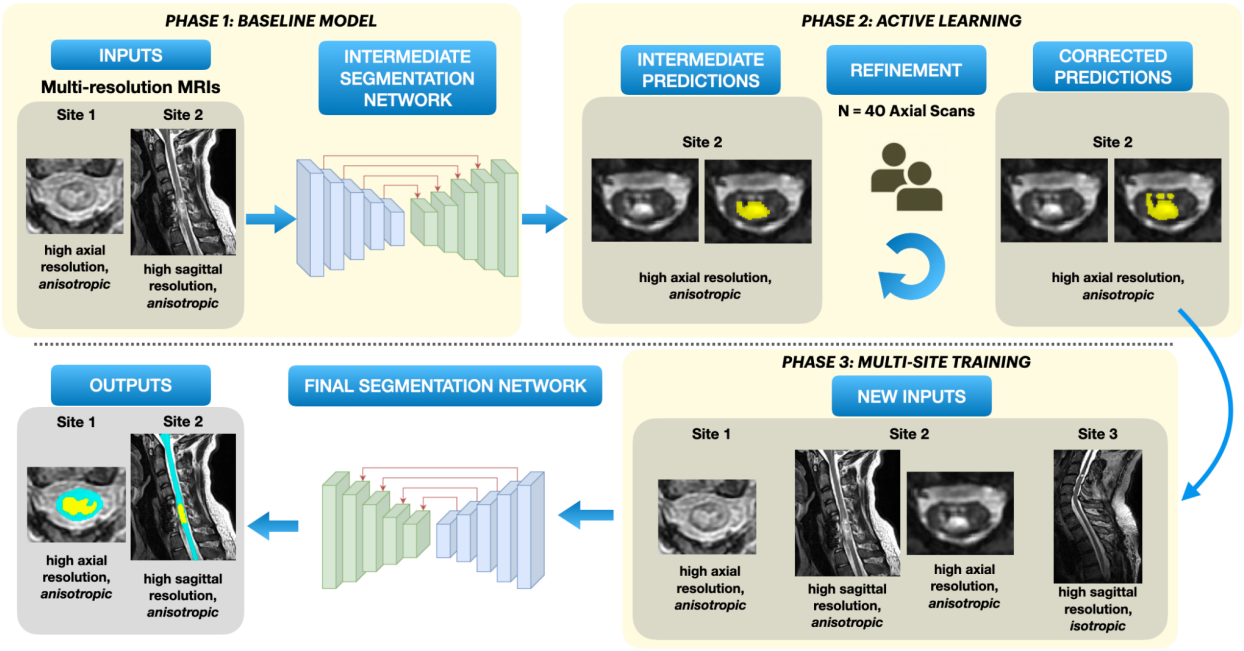

Figure 2: Overview of our segmentation approach. Phase 1: A baseline model is trained on data consisting of T2-weighted images with axial and sagittal orientations from two sites. Phase 2: *Active learning* – Initial batch of automatic predictions on T2-weighted axial images from site 2 are obtained, followed by manual corrections. Phase 3: Along with the newly corrected axial images, isotropic T2-weighted sagittal images from site 3 are added to the original dataset for multi-site training. The final model is trained to segment both spinal cord and lesion simultaneously.

2.4 Evaluation Protocol

We created two independent test sets (site 1: n=19, site 2: n=16), following the 80/20 train/test splitting ratio. We trained five models, each with a different train/test splitting, to avoid biasing the model towards a particular dataset split. The model's performance on the lesion and SC segmentation were evaluated independently within each test set by comparing it with open-source methods available in Spinal Cord Toolbox (SCT) (11): `sct_propseg` (26), `sct_deepseg_sc` (23), and the recently proposed `contrast-agnostic` SC segmentation model (27). Due to the lack of existing state-of-the-art, open-source methods for SCI lesion segmentation, we compared the `SCIseg` 3D model with its 2D version.

2.5 Evaluation Metrics

For quantitative validation, we used the segmentation metrics from the open-source ANIMA toolkit[1]. For SC segmentation, we presented the Dice coefficient and the relative volume error

[1] https://anima.readthedocs.io/en/latest/index.html

(RVE), whereas, for lesion segmentation, we reported the Dice coefficient, average surface distance, lesion-wise positive predictive value (PPV), lesion-wise sensitivity, and F_1 score (28).

2.6 Quantitative MRI Biomarkers

We used the SCT's `sct_analyze_lesion` function to automatically compute the total lesion volume, intramedullary lesion length, and maximal axial damage ratio (6) from the manual ground-truth lesion masks and the automatic predictions using the proposed `SCIseg` 3D model. The quantitative biomarkers were then correlated with the clinical scores.

2.7 Statistical Analysis

Statistical analysis was performed using the SciPy Python library v1.11.4 (29). Data normality was tested using D'Agostino and Pearson's normality test, between-group comparisons were performed using the Wilcoxon signed-rank test, and correlations were examined using the Spearman rank-order correlation.

3. Results

3.1 Patient Characteristics

A total of 191 patients (mean age ± standard deviation 48.1 ± 17.9, 142 males, 42 females, 7 sex not reported) with 231 MRI scans from three sites with different treatment strategies (pre-operative/post-operative) and lesion etiologies (traumatic/ischemic) were included in this study (Figure 1, Table 1). Patients were scanned across scanners from different manufacturers (Siemens, Philips, GE) with different field strengths (1T, 1.5T, 3T). T2-weighted images used in this study had heterogeneous image resolutions, slice thicknesses, and orientations.

Table 1: Characteristics of the Study Cohort

	Site 1	Site 2	Site 3
Number of patients	97	80	14
Number of MRI scans	142	80	14
Sex (male/female)	66/25*	65/15	11/2#
Age (mean ± standard deviation)	51.0 ± 19.1	45.8 ± 16.4	42.9 ± 16.7
Age range	17–83	15–81	21-65
MRI manufacturers	Siemens (n=91), GE (n=5), Philips (n=1)	Siemens (n=20), GE (n=60)	Siemens (n=14)
MRI field strength	3T (n=37), 1.5T (n=59), 1T (n=1)	3T (n=21), 1.5T (n=59)	3T (n=14)
Sequence parameters	SAGITTAL T2-weighted: voxel size 0.34 × 0.34 mm to 0.96 × 0.96 mm; slice thickness 2.2 mm to 4.8 mm AXIAL T2-weighted: voxel size 0.35 × 0.35 mm to 0.78 × 0.78 mm; slice thickness 1.0 mm to 7.0 mm	AXIAL T2-weighted: voxel size 0.31 × 0.31mm to 0.78 × 0.78mm; slice thickness from 3.0 mm to 6.0 mm	ISOTROPIC T2-weighted: voxel size 0.84 × 0.84 × 0.94 mm to 0.875 × 0.875 × 0.9mm

*Sex not reported for 6 patients

#Sex not reported for 1 patient

3.2 Automatic Spinal Cord and Lesion Segmentation in SCI

Table 2 shows the quantitative results of `SCIseg 3D` on test sets of the two sites. We observed that SC segmentations from the model are quite stable across different data splits despite the presence of artifacts in the images. However, for lesion segmentation, the model showed better performance on site 2 compared to site 1 with a high standard deviation across splits.

Table 2. Quantitative performance of the proposed `SCIseg 3D` model. The metrics are averaged across 5 seeds.

Metric	**Spinal Cord Segmentation**		
	Site 1 (n=113)	Site 2 (n=78)	Average (n=191)
Dice Score	0.90 ± 0.07	0.94 ± 0.03	0.92 ± 0.06
RVE %	-1.14 ± 11.19	0.01 ± 8.00	-0.67 ± 10.00
Surface Distance	0.06 ± 0.32	0.00 ± 0.00	0.04 ± 0.25
	Lesion Segmentation		
Dice Score	0.48 ± 0.31	0.76 ± 0.16	0.60 ± 0.29
Surface Distance	1.14 ± 2.79	0.19 ± 0.82	0.72 ± 2.19
Lesion-wise PPV	0.53 ± 0.47	0.89 ± 0.24	0.68 ± 0.43
Lesion-wise Sensitivity	0.73 ± 0.43	0.93 ± 0.20	0.82 ± 0.36
F_1 Score	0.52 ± 0.47	0.89 ± 0.21	0.68 ± 0.42

3.3 Comparison with Other Methods

We compared the SC segmentation performance of our `SCIseg 3D` model with other methods: `sct_propseg`, `sct_deepseg_sc 2D`, `sct_deepseg_sc 3D`, `contrast-agnostic`, and `SCIseg 2D` (Figure 3, Figure 4). The half-violin plots in Figure 4 show the distribution of the Dice scores and RVE for test scans across all seeds and the scatter plots show the performance of the models on each test scan.

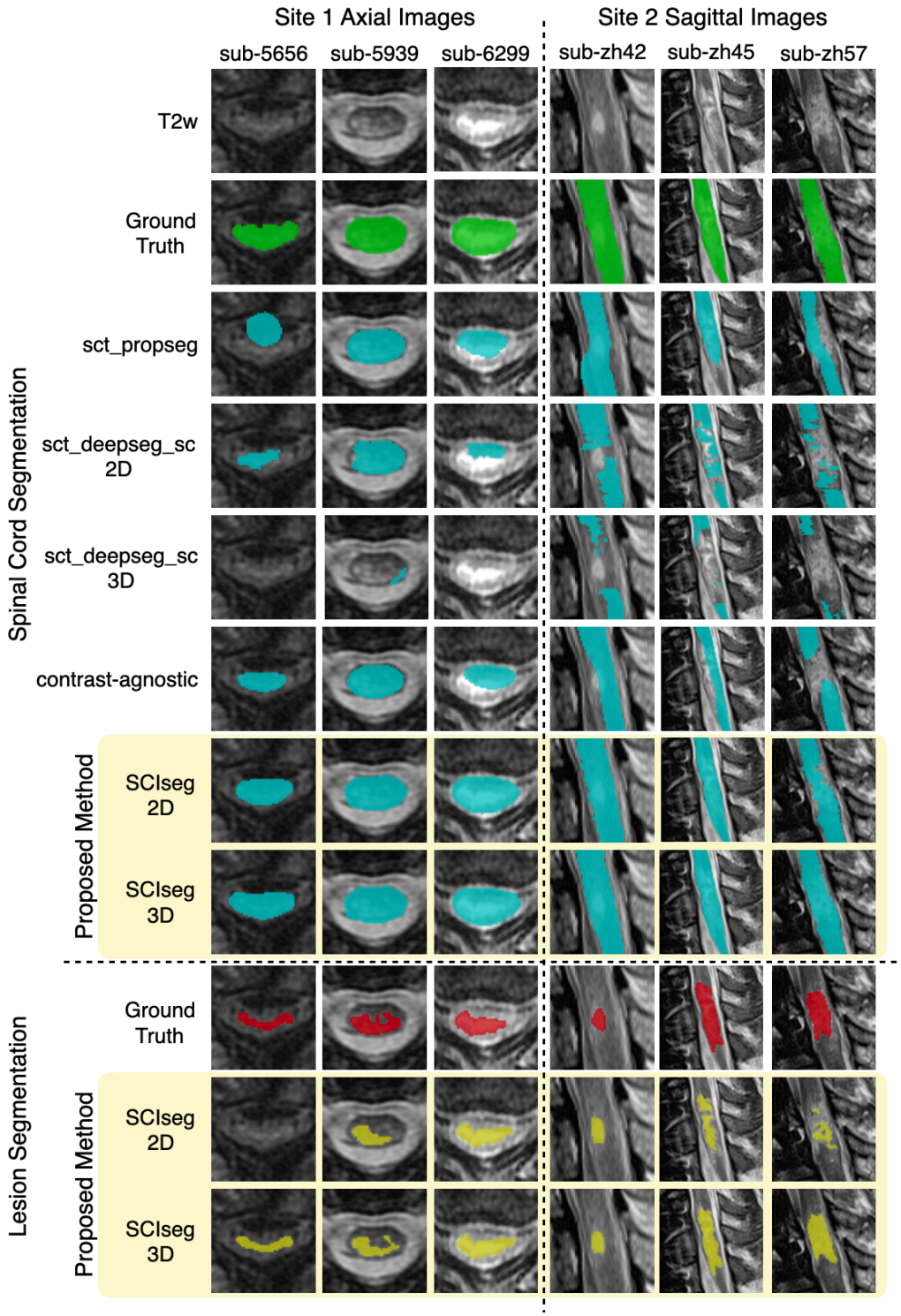

Figure 3: Comparison of `SCIseg` with baseline methods for the spinal cord and lesion segmentation on patients from site 1 and site 2. Notice that `SCIseg` 3D provides the best results qualitatively for both spinal cord and lesion segmentation at the site of lesions.

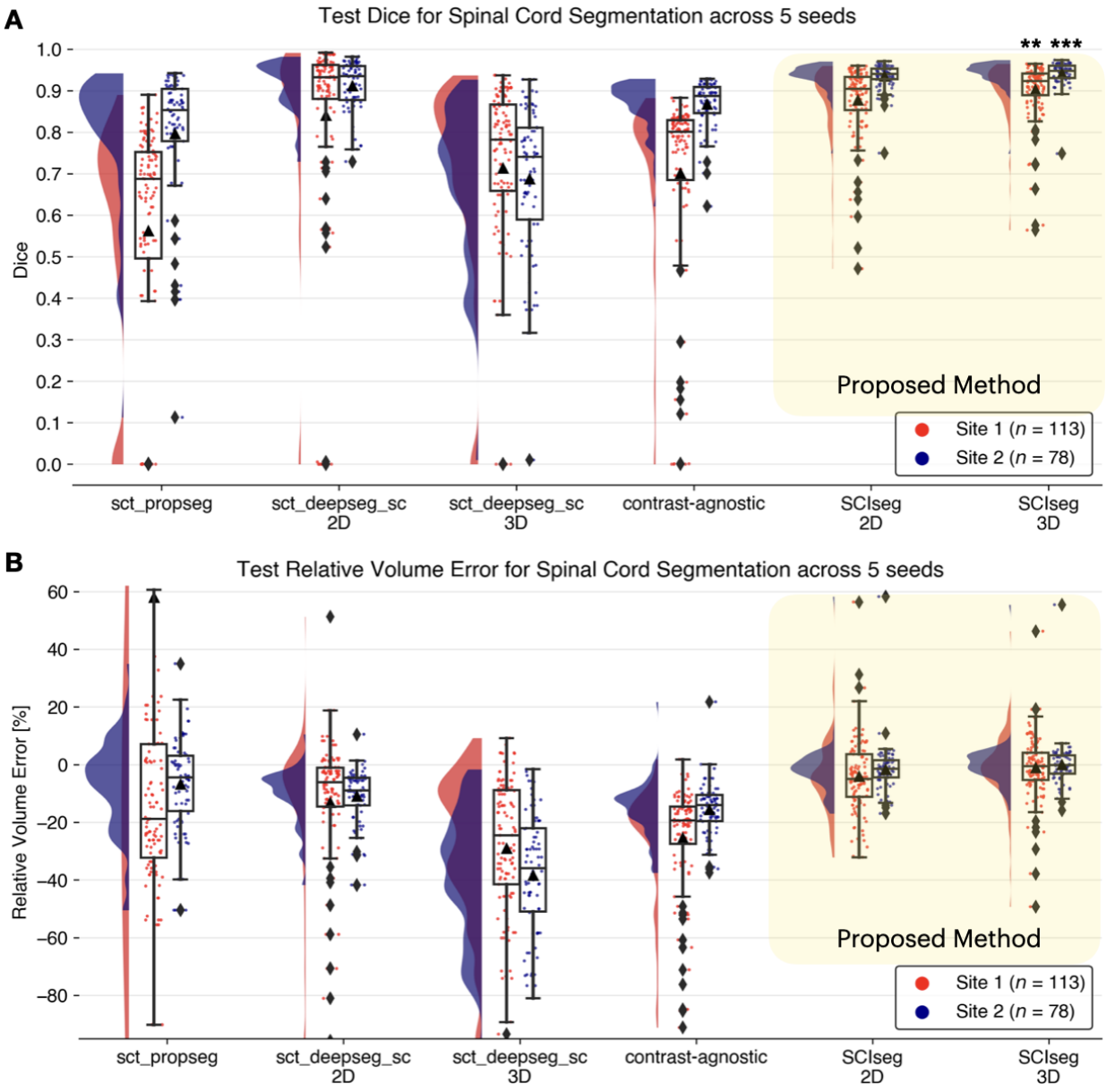

Figure 4: Raincloud plots comparing the (A) Dice scores (best: 1; worst: 0) and (B) relative volume error (in %, best: 0%) across various spinal cord segmentation methods. The numbers in the legend represent the number of test scans in each site summed across 5 different training seeds. Notice that although the `sct_deepseg_sc` 2D and `SCIseg` 3D have similar Dice scores, the former shows a higher under-segmentation (negative relative volume error) compared to the latter. *** $P < .001$ (two-sided Bonferroni-corrected pairwise Wilcoxon signed-rank test for `SCIseg` 3D with all baselines), ** $P < .001$ (statistically significant for all pairs except `SCIseg` 3D and `sct_deepseg_sc` 2D)

3.3.1 Spinal Cord Segmentation

Our model, `SCIseg 3D`, achieves the best segmentation performance across all baselines (Figure 4A). For site 1, we observed relatively more under-/over-segmented predictions shown by a larger spread of scatter points around 0% in the RVE plot (Figure 4B). This can be attributed to the heavy interference caused by metal artifacts in these scans. For site 2, our model performs quite robustly across all test scans. While the performance of `sct_deepseg_sc 2D` is quite similar to that of our model, it must be noted that the ground-truth segmentation masks were originally generated from `sct_deepseg_sc 2D` and then manually corrected. As a result, the Dice scores obtained from this model are inherently biased to be higher than the rest of the methods. However, unlike `SCIseg 3D`, it outputs empty predictions for a few subjects in site 1 (shown by the diamond at Dice=0). More importantly, it must be noted that all the candidate models for comparison were trained specifically for segmenting the spinal cord, whereas `SCIseg 3D` can segment both SC and lesions simultaneously.

3.3.2 Lesion Segmentation

Table 3 presents a comparison between the 2D and 3D variants of the `SCIseg` model. The 3D model performs significantly better than the 2D model across all metrics for both sites. As for the performance within sites, the model's performance on site 2 is higher than that of site 1. Through visual quality control, we noticed that site 1 contained several patients with metal implants causing heavy image artifacts and patients spanned different SCI phases (acute and sub-acute) with various degrees of lesion hyperintensity, thus making automatic segmentation challenging. Despite these issues, the `SCIseg` model provides a good starting point for obtaining lesion segmentations instead of manually annotating lesions from scratch.

Table 3. Lesion segmentation performance of the `SCIseg` models. The metrics are averaged across 5 different training seeds. Bold represents better performance.

Metric	SCIseg 2D		SCIseg 3D	
	Site 1	Site 2	Site 1	Site 2
Dice Score (↑)	0.26 ± 0.30	0.66 ± 0.23	**0.48 ± 0.31** †	**0.76 ± 0.16** †
Surface Distance (↓)	6.42 ± 24.11	0.38 ± 1.18	**1.14 ± 2.79** †	**0.19 ± 0.82** $^{\pi}$
Lesion-wise PPV (↑)	0.22 ± 0.39	0.78 ± 0.35	**0.53 ± 0.47** †	**0.89 ± 0.24** $^{\pi}$
Lesion-wise Sensitivity (↑)	0.43 ± 0.49	0.87 ± 0.31	**0.73 ± 0.43** †	**0.93 ± 0.20** ‡
F_1 Score (↑)	0.23 ± 0.40	0.79 ± 0.34	**0.52 ± 0.47** †	**0.89 ± 0.21** †

Note: Data are means ± standard deviations. The best value for surface distance is 0.0 and 1.0 for the rest of the metrics.

† Statistically significant compared to `SCIseg` 2D. Wilcoxon signed-rank test *(P < .001)*

ᵠ Statistically significant compared to `SCIseg` 2D. Wilcoxon signed-rank test *(P < .01)*

‡ Statistically significant compared to `SCIseg` 2D. Wilcoxon signed-rank test *(P < .05)*

3.4 Effect of Active Learning on Lesion Segmentation

We performed an ablation study comparing the model performance after phase 1 (training on 2 sites) and phase 3 (training on 3 sites after active learning). Figure 5A shows the correlation between manual ground truth and automatic predictions for total lesion volume (top) and intramedullary lesion length (bottom). For both sites, a higher agreement between the manually annotated and automatically derived lesion metrics can be observed for the final model after the third phase of training (i.e., solid lines moving closer to the diagonal identity line). The improvement was statistically significant (Wilcoxon signed-rank test, $p < 0.05$) in estimating the total lesion volume for both sites and only the lesion length for site 1.

Figure 5B shows the performance of our baseline model after phase 1 of training (before active learning) on unseen axial T2-weighted images from site 1. The model tends to under-segment the lesions. However, when trained on more data consisting of axial scans from site 1 and isotropic sagittal scans from site 3 during phase 3 of training, we noticed an overall improvement in the segmentations.

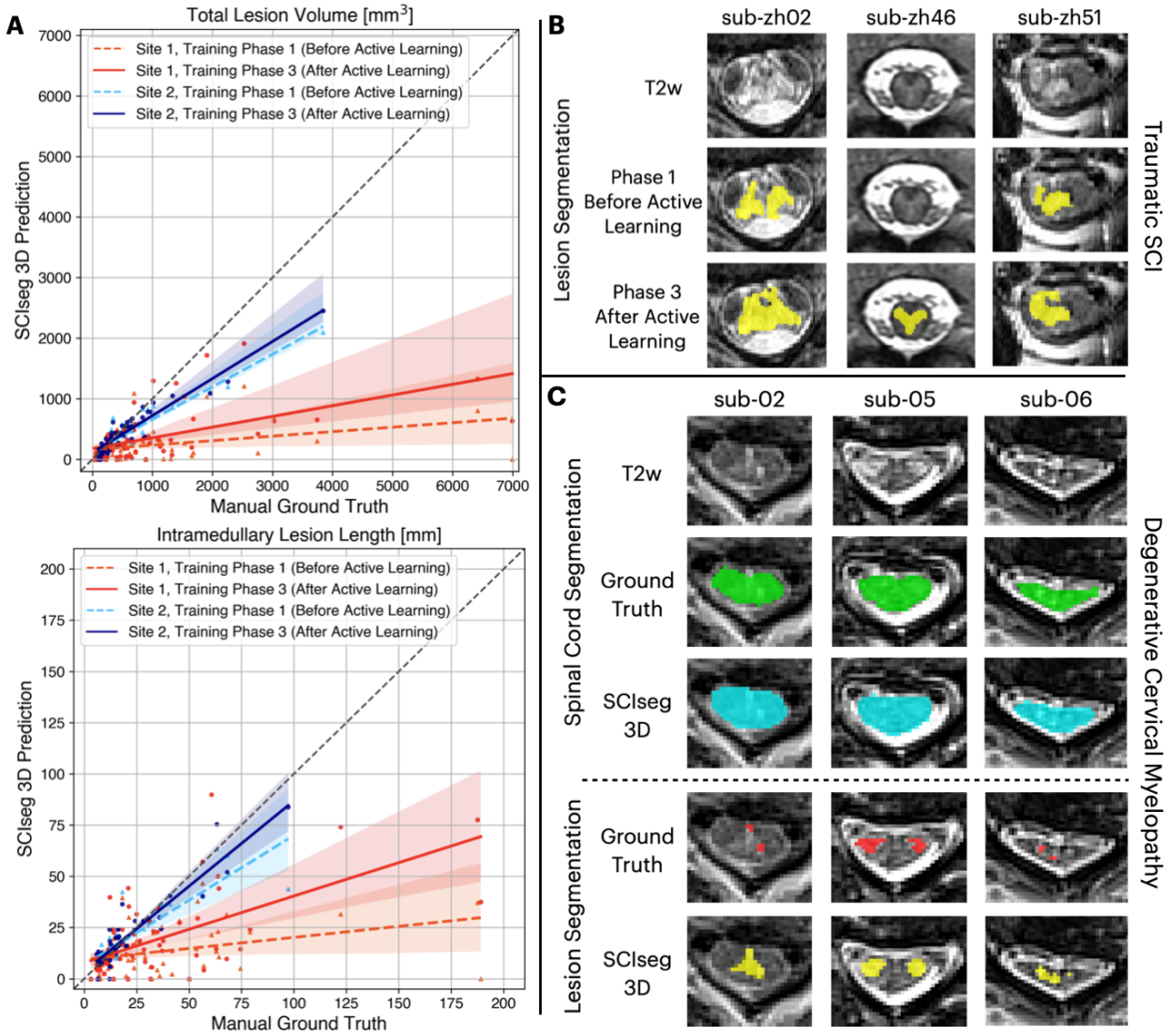

Figure 5: Comparison of model performance before and after active learning. (A) Correlation plots for total lesion volume (top) and intramedullary lesion length (bottom) computed from manual ground truth (GT) lesion masks (x-axis) and lesion predictions from the proposed `SCIseg` 3D model (y-axis). Within each plot, coloured dashed and solid lines show the agreement between manual GT and automatic predictions before and after active learning, respectively, site 1 (red/orange) and site 2 (blue/light-blue). Note that the model's predictions after active learning show a higher agreement with the manual GT for both sites (i.e., solid lines move closer to the diagonal identity line). (B) SCIseg's predictions on unseen axial images from site 2 before and after active learning. (C) Examples of SCIseg's generalization to non-traumatic SCI (i.e., degenerative cervical myelopathy, DCM) patients. Notice that the model obtains an accurate SC segmentation even at the level of severe compression (sub-06).

3.5 Generalization to Degenerative Cervical Myelopathy

As a result of training on heterogeneous data from 3 sites, our `SCIseg` model is sensitive to any hyperintense signal in the image, thus capable of generalizing to non-traumatic SCI, particularly, DCM. In contrast to traumatic SCI lesions, DCM lesions pose a different challenge in that the lesions are typically smaller and are harder to segment in cases of severe SC compression. Figure 5C shows qualitative examples of SC and lesion segmentation on a dataset of DCM patients unseen during training. Interestingly, in cases where the ground-truth lesion masks were under-segmented, the model provided a better and more complete segmentation of the lesion. Furthermore, the SC segmentations are accurate even for slices with severe SC compression (Figure 5C, `sub-06`).

3.6 Correlation between Clinical Scores and MRI Biomarkers

Figure 6 illustrates the relationship between clinical scores (specifically, pinprick, light touch, and lower extremity motor scores) and quantitative MRI biomarkers (namely, lesion volume, lesion length, and maximal axial damage ratio) calculated from both manual ground-truth lesion masks and lesions predicted using SCIseg 3D. We observed a statistically significant correlation between the discharge pinprick score and all lesion biomarkers (see Figure 6, column 1). In the case of discharge lower extremity motor scores, there was a statistically significant correlation with both the total lesion volume and the maximal axial damage ratio (refer to Figure 6, column 3). The Wilcoxon signed-rank test between manual (green) vs. SCIseg-predicted (yellow) lesion biomarkers revealed no statistically significant ($P > .05$) difference. This suggests that the lesions predicted by `SCIseg` are statistically comparable to those identified manually, implying reliability in SCIseg's predictive capabilities.

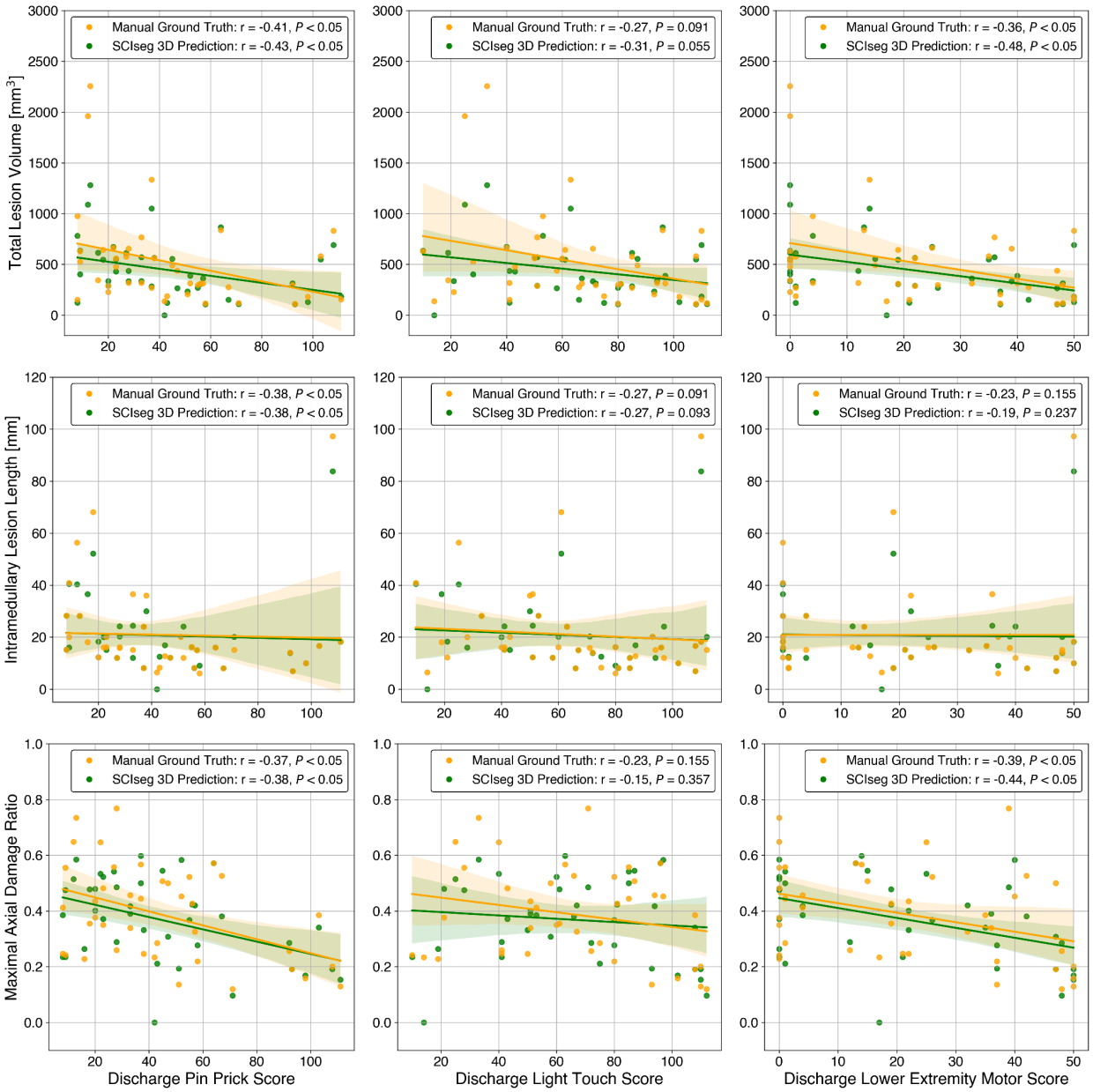

Figure 6: Correlation analyses between discharge clinical scores (x-axis) and quantitative MRI biomarkers (y-axis) for site 2 (Spearman correlation coefficient and p-value are shown in the legends). The rows show quantitative MRI biomarkers and the columns show clinical scores. The quantitative MRI biomarkers were computed from both manual ground-truth lesion masks (yellow) and automatic predictions using `SCIseg` 3D (green). For all plots, no statistically significant difference (Wilcoxon signed-rank test: $P > .05$) was found between quantitative MRI biomarkers derived from manual vs. automatically predicted lesion masks.

4. Discussion

This study introduced a DL-based model, `SCIseg`, for the automatic segmentation of the spinal cord and T2-weighted hyperintense lesions in SCI patients. The model was trained and evaluated on a cohort of 191 SCI patients with 231 scans spanning three sites, different scanner manufacturers, and heterogeneous image resolutions and orientations. To our best knowledge, `SCIseg` is the first open-source, automatic method for lesion and spinal cord segmentation in SCI, performing well across sites, scanners, manufacturers, treatment strategies (pre-/post-operative), and lesion etiologies (traumatic/ischemic). It also generalizes to DCM patients, obtaining accurate segmentations for both lesions and SC at the compression levels. Correlation analyses between clinical scores and MRI-derived biomarkers revealed no statistically significant differences ($P > .05$) when comparing manually defined ground truth and automatically derived lesion masks.

Segmentation of T2-weighted hyperintense lesions in SCI images poses an extremely challenging task mainly due to the evolving appearance of lesions in different injury phases (e.g., acute, sub-acute, intermediate) (1,30). The surgical implants in the postoperative MRI scans also cause severe image artifacts, interfering with automatic image analysis. As a result, researchers often (semi) manually annotate both spinal cord and lesions in SCI patients (6,13,14), which is a time-consuming process susceptible to intra- and inter-rater variability (2). Only a few studies exist in the literature discussing the importance of automatic segmentation in SCI scans (20,31). McCoy et al.'s study (20) is the closest to ours as it presented the first DL method for the segmentation of SC and lesions in SCI. Nevertheless, there are several important distinctions between the two studies. While their model was trained on axial images of pre-operative SCI patients from a single site, our model was trained on multi-site data consisting of pre- and post-operative SCI patients with different image orientations. Moreover, our model was exposed to more heterogeneous data and therefore demonstrated better generalization to both traumatic and non-traumatic lesion etiologies. More importantly, our work is open-source, further enabling reproducible, multi-site studies in SCI.

As the segmentation performance is constrained by the low data quality and small dataset sizes in SCI, we showed that implementing a multi-phase training strategy, which includes (semi-automatic) active learning to progressively expand the dataset size and incorporating diverse data distributions into the training set, contributes significantly to enhancing the model performance. Furthermore, a region-based training strategy that jointly segments the SC and the lesion is more efficient than training two individual models for SC and lesion segmentation, respectively.

This study has a few limitations. First, the model was trained on a single T2-weighted contrast with hyperintense lesions. While the T2-weighted contrast is commonly used in clinics (30,32),

the model could benefit from the complementary information about the lesion characteristics from other contrasts. Second, the lesions were considered as single, independent entities for automatic segmentation. This does not take into account the pathophysiological changes occurring near the site of injury during the acute phase characterized by edema and hemorrhage, which manifest themselves as different MRI signal abnormalities.

There exist several promising avenues for future work. The segmentation models can be improved by using more fine-grained ground-truth masks, where edema and hemorrhage could be treated as separate classes. Training a model on pre-operative traumatic SCI data using these ground-truth masks would have a major impact on improving the initial classification of the disease and further prognostication (14). While the model generalizes reasonably well to DCM lesions, there is a scope for improvement, especially, by adding the DCM cohort to the existing training set or by training a DL model exclusively on DCM data. Previous studies have reported the presence of hyperintense T2-weighted lesions in up to 64% of DCM patients (9,33,34) and explored the relationship between structural and functional damages (35). Such studies would greatly benefit from an automatic DCM lesion segmentation method.

In conclusion, this study presented `SCIseg`, an automatic DL-based method for the segmentation of SC and T2-weighted hyperintense lesions in SCI images. The work has addressed several limitations of previous studies, first, a large retrospective cohort consisting of 191 patients spanning three sites was used, second, MRI data was acquired using scanners from different manufacturers, and third, a single model was trained to segment *both* SC and lesions. More importantly, the methodology has been designed to ensure reproducibility and enable large-scale, reproducible prospective studies. The model is open-source and accessible via SCT. We hope that `SCIseg` will benefit clinicians and patients by providing additional diagnostic and prognostic information, serving as a basis for further studies assessing optimal rehabilitation from a customized patient-based perspective.

Code Availability Statement

To facilitate reproducibility and open science principles, all codes, processing scripts, and results are shared as open-source and freely available to the whole community at https://github.com/ivadomed/model_seg_sci.

Acknowledgements

We thank Nick Guenther and Mathieu Guay-Paquet for their assistance with the management of the datasets, Joshua Newton for his contributions in helping us implement the algorithm to SCT, Maxime Bouthillier for his help in correcting parts of the manuscript, Drs. Thierry Albert, Bertrand Baussart, Caroline Hugeron, Hugues Pascal Moussellard, Frédéric Petit and Marc-Antoine Rousseau for helping with patient recruitment in Paris, Dr. Serge Rossignol and the Multidisciplinary Team on Locomotor Rehabilitation (Regenerative Medicine and Nanomedicine, CIHR), and we thank all patients.

Funding

Funded by the Canada Research Chair in Quantitative Magnetic Resonance Imaging [CRC-2020-00179], the Canadian Institute of Health Research [PJT-190258], the Canada Foundation for Innovation [32454, 34824], the Fonds de Recherche du Québec - Santé [322736, 324636], 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, the Association Française contre les Myopathies (AFM), the Institut pour la Recherche sur la Moelle épinière et l'Encéphale (IRME), and the Ministry of Health of the Czech Republic, grant nr. NU22-04-00024, the National Institutes of Health Eunice Kennedy Shriver National Institute of Child Health and Development (R03HD094577). ACS is supported by the National Institutes of Health – K01HD106928 and R01NS128478 and the Boettcher Foundation's Webb-Waring Biomedical Research Program. KAW is supported by the National Institutes of Health – K23NS104211, L30NS108301, R01NS128478. JV received funding from the European Union's Horizon Europe research and innovation programme under the Marie Sktodowska-Curie grant agreement No 101107932. NKE is supported by the Fonds de Recherche du Québec Nature and Technologie (FRQNT) Doctoral Training Scholarship and in part by the FRQNT Strategic Clusters Program (2020-RS4-265502 - Centre UNIQUE - Union Neurosciences & Artificial Intelligence – Quebec and in part, by funding from the Canada First Research Excellence Fund through the TransMedTech Institute.

References

1. Ahuja CS, Wilson JR, Nori S, et al. Traumatic spinal cord injury. Nat Rev Dis Primers. 2017;3:17018.

2. David G, Mohammadi S, Martin AR, et al. Traumatic and nontraumatic spinal cord injury: pathological insights from neuroimaging. Nat Rev Neurol. 2019;15(12):718–731.

3. Miyanji F, Furlan JC, Aarabi B, Arnold PM, Fehlings MG. Acute cervical traumatic spinal cord injury: MR imaging findings correlated with neurologic outcome--prospective study with 100 consecutive patients. Radiology. Radiological Society of North America (RSNA); 2007;243(3):820–827.

4. Dobran M, Aiudi D, Liverotti V, et al. Prognostic MRI parameters in acute traumatic cervical spinal cord injury. Eur Spine J. 2023;32(5):1584–1590.

5. Huber E, Lachappelle P, Sutter R, Curt A, Freund P. Are midsagittal tissue bridges predictive of outcome after cervical spinal cord injury? Ann Neurol. 2017;81(5):740–748.

6. Smith AC, Albin SR, O'Dell DR, et al. Axial MRI biomarkers of spinal cord damage to predict future walking and motor function: a retrospective study. Spinal Cord. 2021;59(6):693–699.

7. Kurpad S, Martin AR, Tetreault LA, et al. Impact of Baseline Magnetic Resonance Imaging on Neurologic, Functional, and Safety Outcomes in Patients With Acute Traumatic Spinal Cord Injury. Global Spine Journal. SAGE Publications; 2017;7(3 Suppl):151S.

8. Pfyffer D, Huber E, Sutter R, Curt A, Freund P. Tissue bridges predict recovery after traumatic and ischemic thoracic spinal cord injury. Neurology. 2019;93(16):e1550–e1560.

9. Martin AR, De Leener B, Cohen-Adad J, et al. A novel MRI biomarker of spinal cord white matter injury: T2∗-weighted white matter to gray matter signal intensity ratio. AJNR Am J Neuroradiol. 2017;38(6):1266–1273.

10. Badhiwala JH, Ahuja CS, Akbar MA, et al. Degenerative cervical myelopathy - update and future directions. Nat Rev Neurol. 2020;16(2):108–124.

11. De Leener B, Lévy S, Dupont SM, et al. SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data. Neuroimage. 2017;145(Pt A):24–43.

12. Bischof A, Papinutto N, Keshavan A, et al. Spinal Cord Atrophy Predicts Progressive Disease in Relapsing Multiple Sclerosis. Ann Neurol. 2022;91(2):268–281.

13. Smith AC, Weber KA 2nd, O'Dell DR, Parrish TB, Wasielewski M, Elliott JM. Lateral Corticospinal Tract Damage Correlates With Motor Output in Incomplete Spinal Cord Injury. Arch Phys Med Rehabil. 2018;99(4):660–666.

14. Mummaneni N, Burke JF, DiGiorgio AM, et al. Injury volume extracted from MRI predicts

neurologic outcome in acute spinal cord injury: A prospective TRACK-SCI pilot study. J Clin Neurosci. J Clin Neurosci; 2020;82(Pt B):231–236.

15. Vallotton K, Huber E, Sutter R, Curt A, Hupp M, Freund P. Width and neurophysiologic properties of tissue bridges predict recovery after cervical injury. Neurology. 2019;92(24):e2793–e2802.

16. Pfyffer D, Vallotton K, Curt A, Freund P. Predictive Value of Midsagittal Tissue Bridges on Functional Recovery After Spinal Cord Injury. Neurorehabil Neural Repair. 2021;35(1):33–43.

17. Khan O, Badhiwala JH, Wilson JRF, Jiang F, Martin AR, Fehlings MG. Predictive Modeling of Outcomes After Traumatic and Nontraumatic Spinal Cord Injury Using Machine Learning: Review of Current Progress and Future Directions. Neurospine. 2019;16(4):678–685.

18. Dietz N, Vaitheesh Jaganathan, Alkin V, Mettille J, Boakye M, Drazin D. Machine learning in clinical diagnosis, prognostication, and management of acute traumatic spinal cord injury (SCI): A systematic review. J Clin Orthop Trauma. 2022;35:102046.

19. Asgari Taghanaki S, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G. Deep semantic segmentation of natural and medical images: a review. Artificial Intelligence Review. 2021;54(1):137–178.

20. McCoy DB, Dupont SM, Gros C, et al. Convolutional Neural Network-Based Automated Segmentation of the Spinal Cord and Contusion Injury: Deep Learning Biomarker Correlates of Motor Impairment in Acute Spinal Cord Injury. AJNR Am J Neuroradiol. 2019;40(4):737–744.

21. Smith AC, O'Dell DR, Thornton WA, et al. Spinal Cord Tissue Bridges Validation Study: Predictive Relationships With Sensory Scores Following Cervical Spinal Cord Injury. Top Spinal Cord Inj Rehabil. 2022;28(2):111–115.

22. Cohen-Adad J, El Mendili M-M, Lehricy S, et al. Demyelination in the Injured Human Spinal Cord Detected with Diffusion & Magnetization Transfer Imaging. . https://archive.ismrm.org/2011/0398.html.

23. Gros C, De Leener B, Badji A, et al. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. Neuroimage. 2019;184:901–915.

24. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–211.

25. Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med Image Anal. 2021;71:102062.

26. De Leener B, Kadoury S, Cohen-Adad J. Robust, accurate and fast automatic

segmentation of the spinal cord. Neuroimage. 2014; doi: 10.1016/j.neuroimage.2014.04.051.

27. Bédard S, Enamundram NK, Tsagkas C, et al. Towards contrast-agnostic soft segmentation of the spinal cord. arXiv [eess.IV]. 2023. http://arxiv.org/abs/2310.15402.

28. Commowick O, Istace A, Kain M, et al. Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. Sci Rep. 2018;8(1):13650.

29. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. Nature Publishing Group; 2020;17(3):261–272.

30. Freund P, Seif M, Weiskopf N, et al. MRI in traumatic spinal cord injury: from clinical assessment to neuroimaging biomarkers. Lancet Neurol. 2019;18(12):1123–1135.

31. Blanc C, Shahrampour S, Mohamed FB, de Leener B. Combining PropSeg and a convolutional neural network for automatic spinal cord segmentation in pediatric populations and patients with spinal cord injury. Int J Imaging Syst Technol. John Wiley & Sons, Ltd; 2023;n/a(n/a). doi: 10.1002/ima.22859.

32. Bozzo A, Marcoux J, Radhakrishna M, Pelletier J, Goulet B. The role of magnetic resonance imaging in the management of acute spinal cord injury. J Neurotrauma. 2011;28(8):1401–1411.

33. Nouri A, Martin AR, Kato S, Reihani-Kermani H, Riehm LE, Fehlings MG. The Relationship between MRI Signal Intensity Changes, Clinical Presentation, and Surgical Outcome in Degenerative Cervical Myelopathy. Spine . 2017;42(24):1851–1858.

34. Martin AR, Tadokoro N, Tetreault L, et al. Imaging Evaluation of Degenerative Cervical Myelopathy: Current State of the Art and Future Directions. Neurosurgery Clinics of North America. Neurosurg Clin N Am; 2018. p. 33–45. doi: 10.1016/j.nec.2017.09.003.

35. Scheuren PS, David G, Kipling Kramer JL, et al. Combined Neurophysiologic and Neuroimaging Approach to Reveal the Structure-Function Paradox in Cervical Myelopathy. Neurology. 2021; doi: 10.1212/WNL.0000000000012643.