



Monitoring morphometric drift in lifelong learning segmentation of the spinal cord

Enamundram Naga Karthik^{1,2}, Sandrine Bédard¹, Jan Valošek^{1,2,3,4}, Christoph S. Aigner^{5,6}, Elise Bannier⁷, Josef Bednařík^{8,9}, Virginie Callot^{10,11}, Anna Combes^{12,13}, Armin Curt¹⁴, Gergely David^{14,15}, Falk Eippert¹⁶, Lynn Farner¹⁴, Michael G Fehlings^{17,18}, Patrick Freund^{14,19,20}, Tobias Granberg^{21,22}, Cristina Granziera²³, Ulrike Horn¹⁶, Tomáš Horák^{8,9}, Suzanne Humphreys²⁴, Markus Hupp¹⁴, Anne Kerbrat^{25,26}, Nawal Kinany^{27,28}, Shannon Kolind²⁹, Petr Kudlička^{9,30}, Anna Le Bret¹⁴, Lisa Eunyoung Lee³¹, Caterina Mainero³², Allan R. Martin³³, Megan McGrath¹³, Govind Nair³⁴, Kristin P. O'Grady¹³, Jiwon Oh³⁵, Russell Ouellette^{21,22}, Nikolai Pfender¹⁴, Dario Pfyffer^{14,36}, Pierre-François Pradat³⁷, Alexandre Prat³⁸, Emanuele Pravata^{39,40}, Daniel S. Reich⁴¹, Iliaria Ricchi^{27,28}, Naama Rotem-Kohavi²⁴, Simon Schading-Sassenhausen¹⁴, Maryam Seif^{14,19}, Andrew Smith⁴², Seth A Smith¹³, Grace Sweeney¹³, Roger Tam⁴³, Anthony Traboulsee⁴⁴, Constantina Andrada Treaba³², Charidimos Tsagkas^{41,23}, Zachary Vavasour⁴³, Dimitri Van De Ville^{27,28}, Kenneth Arnold Weber II⁴⁵, Sarath Chandar^{2,46}, Julien Cohen-Adad^{1,2,47,48}

¹NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada

²Mila—Quebec AI Institute, Montreal, QC, Canada

³Department of Neurosurgery, Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc, Czechia

⁴Department of Neurology, Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc, Czechia

⁵Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany

⁶Max Planck Research Group MR Physics, Max Planck Institute for Human Development, Berlin, Germany

⁷Department of Neuroradiology, Rennes University Hospital, Rennes, France

⁸Department of Neurology, University Hospital Brno, Brno, Czechia

⁹Faculty of Medicine, Masaryk University, Brno, Czechia

¹⁰Aix-Marseille Univ, CNRS, CRMBM, Marseille, France

¹¹APHM, CHU Timone, CEMEREM, Marseille, France

¹²NMR Research Unit, Queen Square Multiple Sclerosis Centre, UCL Queen Square Institute of Neurology, University College London, London, United Kingdom

¹³Vanderbilt University Institute of Imaging Science, Vanderbilt University Medical Center, Nashville, TN, United States

¹⁴Spinal Cord Injury Center, Balgrist University Hospital, University of Zurich, Zurich, Switzerland

¹⁵Department of Neuro-Urology, Balgrist University Hospital, University of Zurich, Zurich, Switzerland

¹⁶Max Planck Research Group Pain Perception, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

¹⁷Division of Neurosurgery and Spine Program, Department of Surgery, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

¹⁸Division of Neurosurgery, Krembil Neuroscience Centre, University Health Network, Toronto, ON, Canada

¹⁹Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

²⁰Wellcome Trust Centre for Neuroimaging, Queen Square Institute of Neurology, University College London, London, United Kingdom

²¹Department of Neuroradiology, Karolinska University Hospital, Stockholm, Sweden

²²Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

²³Translational Imaging in Neurology (ThINK), Department of Biomedical Engineering, Faculty of Medicine, Basel, Switzerland

²⁴Praxis Spinal Cord Institute, Vancouver, BC, Canada

²⁵EMPENN Research Team, IRISA, CNRS-INSERM-INRIA, Rennes Université, Rennes, France

²⁶Neurology Department, Rennes University Hospital, Rennes, France

²⁷Neuro-X Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Geneva, Switzerland

²⁸Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

²⁹Division of Neurology, Department of Medicine and the Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, BC, Canada

³⁰Multimodal and Functional Imaging Laboratory, Central European Institute of Technology, Brno, Czechia

³¹Institute of Medical Science, University of Toronto, Toronto, ON, Canada

³²Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, USA; Harvard Medical School, Boston, MA, United States

³³Department of Neurosurgery, University of California Davis, Davis, CA, United States

Received: 14 April 2025 Revision: 20 October 2025 Accepted: 17 December 2025 Available Online: 5 January 2026



³⁴qMRI Core Facility, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, United States

³⁵Barlo MS Centre, Division of Neurology, Department of Medicine, St. Michael's Hospital, Toronto, Canada

³⁶Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Palo Alto, CA, United States

³⁷Department of Neurology, Pitie-Salpetriere Hospital, Paris, France

³⁸Department of Neuroscience, Université de Montréal, Montréal, QC, Canada

³⁹Department of Neuroradiology, Neurocenter of Southern Switzerland, Lugano, Switzerland

⁴⁰Department of Neuroscience, Imaging and Clinical Sciences, Università G. d'Annunzio, Chieti, Italy

⁴¹Translational Neuroradiology Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, United States

⁴²Department of Physical Medicine and Rehabilitation, University of Colorado School of Medicine, Aurora, CO, United States

⁴³School of Biomedical Engineering, Department of Radiology, The University of British Columbia, Vancouver, BC, Canada

⁴⁴Department of Medicine, Division of Neurology, University of British Columbia, BC, Canada

⁴⁵Division of Pain Medicine, Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Palo Alto, CA, United States

⁴⁶Canada CIFAR AI Chair, Toronto, ON, Canada

⁴⁷Functional Neuroimaging Unit, CRIUGM, Université de Montréal, Montreal, QC, Canada

⁴⁸Centre de Recherche du CHU Sainte-Justine, Université de Montréal, Montréal, QC, Canada

Corresponding Authors: Enamundram Naga Karthik (naga-karthik.enamundram@polymtl.ca), Julien Cohen-Adad (julien.cohen-adad@polymtl.ca)

ABSTRACT

Morphometric measures derived from spinal cord segmentations can serve as diagnostic and prognostic biomarkers in neurological diseases and injuries affecting the spinal cord. For instance, the spinal cord cross-sectional area can be used to monitor cord atrophy in multiple sclerosis and to characterize compression in degenerative cervical myelopathy. While robust, automatic segmentation methods to a wide variety of contrasts and pathologies have been developed over the past few years, whether their predictions are stable as the model is updated using new datasets has not been assessed. This is particularly important for deriving normative values from healthy participants. In this study, we present a spinal cord segmentation model trained on a multisite ($n = 75$ sites, 1,631 participants) dataset, including 9 different MRI contrasts and several spinal cord pathologies. We also introduce a lifelong learning framework to automatically monitor the morphometric drift as the model is updated using additional datasets. The framework is triggered by an automatic GitHub Actions workflow every time a new model is created, recording the morphometric values derived from the model's predictions over time. As a *real-world* application of the proposed framework, we employed the spinal cord segmentation model to update a recently introduced normative database of healthy participants containing commonly used measures of spinal cord morphometry. Results showed that (i) our model performs well compared with its previous versions and existing pathology-specific models on the lumbar spinal cord, images with severe compression, and in the presence of intramedullary lesions and/or atrophy achieving an average Dice score of 0.95 ± 0.03 ; (ii) the automatic workflow for monitoring morphometric drift provides a quick feedback loop for developing future segmentation models; and (iii) the scaling factor required to update the database of morphometric measures is nearly constant among slices across the given vertebral levels, showing minimum drift between the current and previous versions of the model monitored by the framework. The code and model are open source and accessible via Spinal Cord Toolbox v7.0.

Keywords: segmentation, MRI, spinal cord, MLOps, lifelong learning, morphometric drift

1. INTRODUCTION

Spinal cord segmentation is relevant for quantifying morphometric changes, such as cord atrophy in multiple sclerosis (MS) (Bautin & Cohen-Adad, 2021; Losseff et al., 1996; Lukas et al., 2013), compression severity in degenerative cervical myelopathy (DCM) (Horáková et al., 2022; Martin et al., 2018), and spared tissue in spinal cord injury (SCI) (Karthik et al., 2024). The development of a robust and accurate spinal cord segmentation tool requires a large sample size which often involves the collaboration of multiple sites and the inclusion of a wide

spectrum of MRI scans spanning various spinal cord pathologies, image resolutions, orientations, contrasts, and potential image artifacts. Consequently, obtaining stable morphometric measurements is challenging, as MRI contrasts with different resolutions (and degrees of anisotropy) have varying levels of partial volume effects, leading to subtle shifts in the boundary between the cord and the cerebrospinal fluid (CSF) (Cohen-Adad et al., 2021a; Valošek & Cohen-Adad, 2024). Furthermore, the stability of morphometric measurements is inherently dependent on the version of the segmentation tool and

may drift as newer versions are released. This poses a challenge in studies where morphometric measures (e.g., cross-sectional area) are monitored across time.

Previous work in automatic spinal cord segmentation has been limited by a lack of standardization, with models often developed in isolation using different procedures for creating ground-truth masks, different model architectures, and varying training strategies (M. Chen et al., 2013; De Leener et al., 2014; Gros et al., 2019; Masse-Gignac et al., 2023; Naga Karthik, Valošek, et al., 2025; Nozawa et al., 2023; Tsagkas et al., 2023). Gros et al. (2019) proposed a collection of contrast-specific models (`sct_deepseg_sc`) trained on healthy controls and MS patients. It uses a convolutional network with 2D kernels, which fails to capture the full spatial context in 3D, resulting in poor performance in DCM and SCI patients with lesions. Masse-Gignac et al. (2023) developed a cascade of two CNNs, trained separately on axial and sagittal T2w scans, for segmenting injured spinal cords, adapting GT masks from `sct_deepseg_sc` 2D. Nozawa et al. (2023) focused on the segmentation of compressed spinal cords with 2D UNets using transfer learning from DeepLabv3 models (L.-C. Chen et al., 2017). Bédard et al. (2025) introduced `contrast_agnostic`, a 3D model trained on a dataset of healthy participants (Cohen-Adad et al., 2021b), which generalizes across contrasts but struggles to segment pathological cases. The existence of numerous specialized models highlights the lack of standardization in the development of an automatic segmentation pipeline and no continuous learning pipeline exists to monitor or mitigate drift in the segmentation performance of these models over time.

Morphometric measures derived from spinal cord segmentations are highly dependent on the method used (Bédard et al., 2025; Cohen-Adad et al., 2021b) and may drift as the methods evolve. This can lead to inconsistencies in normative values across methods. Moreover, morphometric measures exhibit substantial inter-participant variability driven by factors such as age and sex, which limits sensitivity to subtle changes (Bédard et al., 2024; Bédard & Cohen-Adad, 2022; Labounek et al., 2024; Papinutto et al., 2020; Taso et al., 2016; Valošek et al., 2024). One approach to mitigate this variability is to compare them with morphometrics obtained from healthy controls (Bédard et al., 2024; Horáková et al., 2022; Kato et al., 2012; Labounek et al., 2024; Valošek et al., 2024). These normalization techniques assume that the morphometrics from new participants are computed using the same method as the normative reference (Valošek et al., 2024)—an assumption that no longer holds as segmentation methods are iteratively improved upon, highlighting the need for population databases to evolve alongside segmentation techniques.

Given that the aforementioned tools only target a limited set of pathologies, often with few MRI contrasts, there is great value in unifying their specialized analyses into a single model which could work with a substantially larger, cumulative, training set. With segmentation frameworks such as nnUNetV2 (Isensee et al., 2021), which has been widely adopted by the medical imaging community due to its robustness and generalization to several modalities and neural network architectures (Isensee et al., 2024), achieving this objective is now possible. In addition, a standardized training strategy to continuously update models over time, monitor performance drift between various model updates, and manage model retraining would streamline these approaches substantially. Such a lifelong learning framework (Agirre et al., 2021; Liu & Mazumder, 2021; Prapas et al., 2021) ensures that the model remains robust to shifts in the data distribution and continually refine their segmentation performance across the diverse set of contrasts and pathologies (Karthik et al., 2022).

To address these challenges, our study contributes the following:

1. An automatic spinal cord segmentation model trained on a multi-site dataset gathered from 75 sites worldwide. This dataset consisted of nine different MRI contrasts spanning a wide range of image resolutions, including pathologies such as MS (with different phenotypes), traumatic SCI (acute and chronic), and non-traumatic SCI (DCM and ischemic SCI).
2. A lifelong learning framework for developing models to segment new contrasts and pathologies over time. This framework also presents an automatic workflow capable of monitoring the drift in the spinal cord morphometrics across various versions of the models using GitHub Actions.
3. Validation of the lifelong learning framework to update a normative database of spinal cord morphometric measures (Valošek et al., 2024).

The proposed spinal cord segmentation model and normative database are open source and integrated into the Spinal Cord Toolbox (SCT) (De Leener et al., 2017), accessible as of v7.0.

2. MATERIALS AND METHODS

2.1. Data curation

2.1.1. Data and participants

Our “real-world” dataset contains data from 75 sites and 1,631 participants, including healthy participants ($n = 428$), people with degenerative cervical myelopathy (DCM;

$n = 359$), spinal cord injury (SCI; $n = 286$), MS or suspected MS ($n = 164$), amyotrophic lateral sclerosis (ALS; $n = 13$), neuromyelitis optica spectrum disorder (NMOSD; $n = 10$), and syringomyelia (SYR; $n = 1$). The MS cohort spanned different phenotypes, ranging from preclinical MS stage (i.e., radiologically isolated syndrome, RIS; $n = 61$) to clinically definite MS, including relapsing-remitting MS (RRMS; $n = 249$), and primary progressive MS (PPMS; $n = 60$). Within the SCI cohort, the images spanned various phases and lesion etiologies of the injury, namely traumatic ($n = 171$; intermediate and chronic), acute traumatic (pre-decompression) SCI ($n = 95$), ischemic ($n = 13$), hemorrhagic ($n = 5$), and unknown ($n = 2$) lesions. A single participant may contribute one or more different sequences, depending on the site, resulting in a total of 3,453 images (3D volumes¹). The study included 9 different contrasts, namely, T1-weighted (T1w; $n_{vol.} = 318$), T2-weighted (T2w; $n_{vol.} = 1377$), T2*-weighted (T2*w; $n_{vol.} = 499$), diffusion-weighted (DWI; $n_{vol.} = 243$), gradient-echo sequence with (MT-on; $n_{vol.} = 248$) and without (GRE-T1w; $n_{vol.} = 243$) magnetization transfer pulse, phase-sensitive inversion recovery (PSIR; $n_{vol.} = 333$), short tau inversion recovery (STIR; $n_{vol.} = 89$), and MP2RAGE UNIT1 ($n_{vol.} = 103$). The images could cover any of the cervical, thoracic, and lumbar spinal regions (i.e., the model was trained on chunks containing either of those regions). Whole-spine scans covering all regions are not used for training. Spatial resolutions included isotropic (0.8–1 mm), anisotropic axially oriented (in-plane resolution: 0.29–1 mm; slice thickness: 1–9.3 mm), and sagittally oriented (in-plane resolution: 0.28–1 mm; slice thickness: 0.8–4.83 mm) images. Images were acquired at 1T, 1.5T, 3T, and 7T on various scanner manufacturers (Siemens, Philips, and GE). [Figure 1](#) shows the overall summary of the dataset and Supplementary Table S1 provides more details on the distribution of image resolutions for each contrast.

2.1.2. Generating ground-truth masks

We used the GT masks in the spine-generic multi-subject database, generated using the same preprocessing procedure from our previous work ([Bédard et al., 2025](#)). For the newly obtained datasets, we initially performed a quality control (QC) using `sct_qc`, SCT’s visual QC tool ([Valošek & Cohen-Adad, 2024](#)). Four experienced raters (E.N.K., S.B., J.V., J.C.A.) qualitatively assessed the image–GT pairs and flagged images with motion artifacts and poor signal quality to be excluded from training. In cases where the GT masks were under- or over-segmented (e.g., due to the lower contrast at the spinal

cord–cerebrospinal fluid boundary or due to the presence of cord compression), the GT masks were recreated using a combination of the contrast-agnostic model ([Bédard et al., 2025](#)) and manual corrections. In datasets with severe deformations to the spinal cord anatomy (e.g., SCI and DCM), a pathology-specific model, SCISeg ([Karthik et al., 2024](#); [Naga Karthik, Valošek, et al., 2025](#)), was used instead, followed by manual corrections by J.V. and E.N.K. In pathologies involving intramedullary lesions (e.g., MS, SCI, and DCM), lesions were considered part of the spinal cord and included in the GT masks. Manual intervention was required in approximately 9% of the scans (~311 images out of 3,453 in total). All GT masks were binarized using a threshold of 0.5 prior to preprocessing and training to ensure uniformity.

For each site, the data were split participant-wise following an 80–20% train–test split ratio, ensuring that participants with multiple scans (or multiple sessions) were included either in the training set or in the testing set (mutually exclusive). This ensures that no data leakage between train and test splits could occur. After pooling the training and testing data from each participant and each site, the aggregated dataset included 2,945 training and 508 testing images.

2.2. Training protocol

2.2.1. Preprocessing and data augmentation

We chose the nnUNet framework for training our spinal cord segmentation model as it easily allows future retraining of the model with new contrasts and pathologies and can also be readily integrated into existing open-source packages such as SCT ([De Leener et al., 2017](#); [SlicerNNUnet: 3D Slicer nnUNet Integration to Streamline Usage for nnUNet Based AI Extensions, n.d.](#)), facilitating broader use by the spinal cord imaging community.

All images and GT masks were re-oriented to right-posterior-inferior (RPI). The median resolution of images in the training set was $[0.9 \times 0.7 \times 1] \text{ mm}^3$ and the median shape was $[96 \times 320 \times 318]$. Images were resampled to the median resolution using spline interpolation (order = 3), and GT masks were resampled using linear interpolation (order = 1). The patch size was set to $[64 \times 224 \times 160]$. Standard data augmentation transforms in the nnUNet pipeline, being randomly applied, were predefined with a probability (p) and called in the following order: affine transformation (rotation and scaling; $p = 0.2$), Gaussian noise addition ($p = 0.1$), Gaussian smoothing ($p = 0.2$), image brightness augmentation ($p = 0.15$), simulation of low resolution with downsampling and upsampling factors sampled uniformly from $[0.5, 1.0]$ ($p = 0.25$), gamma correction ($p = 0.1$), mirroring transform across all axes.

¹Note that “images” and “volumes” are used interchangeably, both referring to 3D MRI scans.

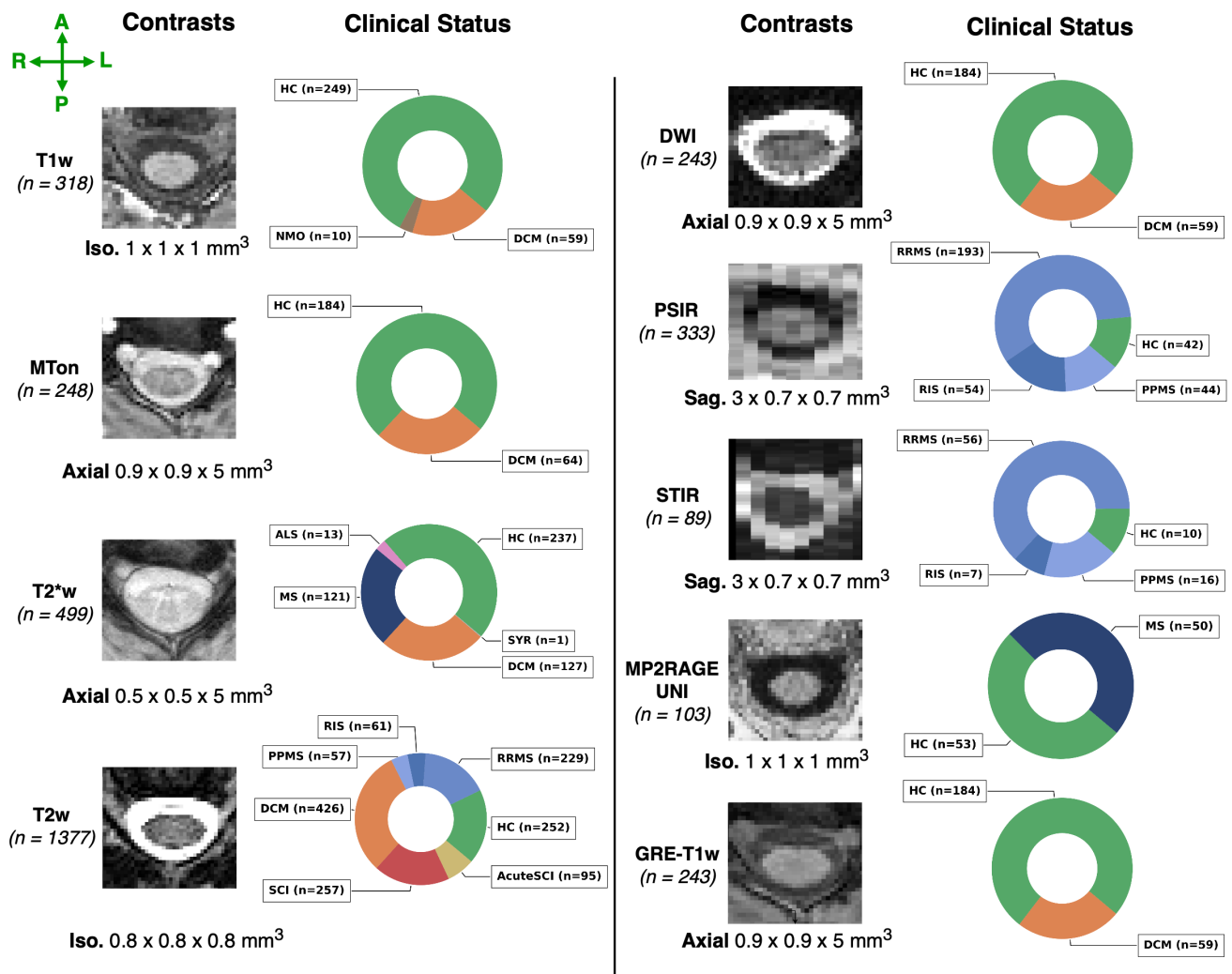


Fig. 1. Overview of the dataset and image characteristics. Representative axial slices of nine contrasts and the total of images used for each contrast in brackets, the orientation (axial/sagittal) along with the median resolution of images. The respective doughnut chart illustrates the proportion of clinical status among the scanned participants, including healthy controls (HC), patients with radiologically isolated syndrome (RIS), patients with multiple sclerosis (MS), and their different phenotypes, including primary progressive (PPMS) and relapsing-remitting (RRMS), patients with amyotrophic lateral sclerosis (ALS), patients with neuromyelitis optica spectrum disorder (NMOSD), pre-decompression acute traumatic SCI (AcuteSCI), post-decompression traumatic spinal cord injury (SCI), degenerative cervical myelopathy (DCM), and syringomyelia (SYR; not shown). Labels indicate the phenotype associated with the patient, with their respective colors shared across contrast sets.

The versatility of these transforms is evident from nnUNet’s success on a diverse range of segmentation tasks. Hence, we decided to keep the predefined set of augmentation transforms. Part of the decision for transformations to be applied also stems from empirical evidence from our prior work (Bédard et al., 2025). Lastly, all images were normalized using z-score normalization.

2.2.2. Network architecture and training hyperparameters

The network architecture is based on the U-Net with 6 layers in the encoder, starting with 32 feature maps at the

initial layer and ending with 320 feature maps at the bottleneck (i.e., $32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 320 \rightarrow 320$) and 5 layers in the decoder with skip connections at each successive resolution. Each layer in the encoder and the decoder consists of a stack of two convolutional blocks, where each block contains a series of convolutional (with $3 \times 3 \times 3$ kernels), instance normalization, and LeakyReLU activation layers (with slope = 0.1). No dropout layers were used. At each layer in the decoder, deep supervision (Dou et al., 2017) was also used, where auxiliary losses from the feature maps at each upsampling resolution are added to the final loss. The network was trained with a combination of Dice (Milletari et al., 2016) and

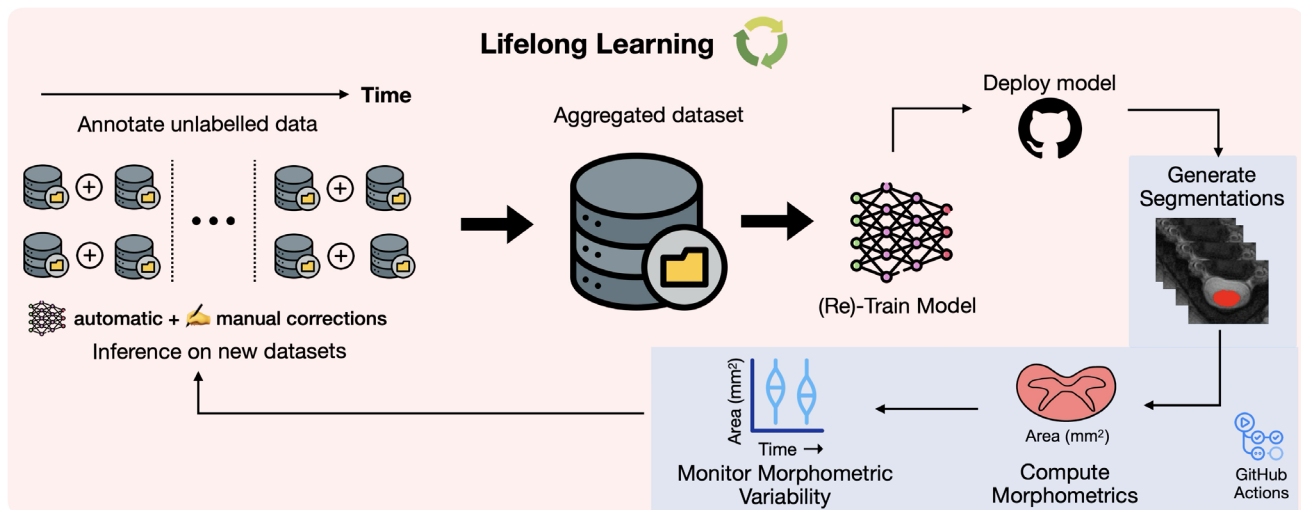


Fig. 2. Overview of the lifelong learning strategy for continuous training of spinal cord segmentation models. Unlabeled data containing various contrasts and pathologies, gathered from multiple sites worldwide, are segmented automatically with an existing state-of-the-art model and undergo visual quality control for inconsistencies in segmentations, excluding data with artifacts. Labeled datasets are aggregated to train the spinal cord segmentation model. Post-training, the model is deployed as an official release, triggering an automatic GitHub Actions workflow that generates the segmentations, computes the morphometrics, and actively monitors the drift in the morphometric variability between the current version of the model and the previously released versions (automated tasks shown in the blue box). As new data arrive, the process is repeated, enabling continuous (re)training of the models to segment a diverse set of contrasts and pathologies.

cross-entropy losses with equal weights. The model was trained using 5-fold cross-validation for 1,000 epochs, a batch size of 2, and with the stochastic gradient descent (SGD) optimizer and a polynomial learning rate scheduler. Checkpoints were saved based on the exponential moving average of Dice score on the validation set computed at the end of each epoch. Supplementary Table S3 contains the values for each hyperparameter mentioned above. All experiments were run on a single 48GB NVIDIA A6000 GPU.

2.3. Lifelong learning for morphometric drift monitoring

We take an *MLOps* (Alla & Adari, 2020; Tabassam, 2023; Treveil et al., 2020) approach to propose our lifelong learning framework for monitoring morphometric drift across various versions of the model (Fig. 2). Once the segmentation model is trained, we deploy the model as an official release on GitHub.² The release triggers an automatic GitHub Actions workflow that (i) downloads the publicly available dataset, (ii) runs the morphometric analysis, (iii) generates the plots quantifying the drift in the performance between the current and previous versions of the model, and (iv) updates the GitHub release

assets by uploading the plots and the morphometric values. It is worth emphasizing that all the above steps are performed automatically once a model is released, thus facilitating model development through continuous integration and continuous deployment (CI/CD) (see Fig. 3 for pseudocode of the workflow). A key element in the automated workflow is a public test set, which must be “frozen” to ensure newer models are evaluated fairly against earlier versions. Accordingly, we used the test set from our previous study³ (Bédard et al., 2025), consisting of 49 healthy participants with 6 contrasts each (T1w, T2w, T2*w, DWI, MT-on, GRE-T1w). Using these data, we computed the spinal cord cross-sectional area (CSA) to monitor morphometric variability. More importantly, monitoring performance drift among models on publicly available participant data avoids data privacy issues when running the morphometric analysis on the cloud using GitHub Actions workflows. Furthermore, running this task after each model finishes training ensures that the deployed model does not drift too much from the stable version (Bédard et al., 2025). We can then use the current version of the model (which is now the new state-of-the-art) to annotate existing or new unlabeled datasets (arriving in the future), perform QC, add them to growing collection of datasets, and

²<https://github.com/sct-pipeline/contrast-agnostic-softseg-spinalcord/releases/tag/v3.0>

³<https://github.com/sct-pipeline/contrast-agnostic-softseg-spinalcord/releases/tag/v2.0>

Algorithm 1 Pseudocode for Monitoring Morphometric Drift

```

name: Run morphometric analysis
on:
  release:
    types: [published]
jobs:
  # job 1: Download the dataset hosted on git-annex
  download_dataset: # define name for the job
    steps:
      # steps performed in the job
      - name: Install git-annex
      - name: Download test data using git-annex
      - name: Cache downloaded dataset

  # job 2: Compute morphometrics
  compute_csa:
    needs: download_dataset # requires previous job to finish
    steps:
      - name: Restore cached dataset
      - name: Install Spinal Cord Toolbox
      - name: Run morphometric analysis on test subset

  # job 3: Generate plots
  generate_plots:
    needs: compute_csa
    steps:
      - name: Generate morphometric drift plots
      - name: Upload plots to GitHub release

```

Fig. 3. Pseudocode of the automatic workflow for monitoring morphometric drift after deploying the segmentation model. The workflow is divided into three jobs: (1) downloading the dataset from `git-annex`, (2) running morphometric analysis (computing CSA) across the test set, and (3) generating plots to monitor drift in morphometric variability and updating the GitHub release with the plots. Note that job 2 is parallelized across several GitHub runners on the cloud, where each runner processes a subset of the test set for computational efficiency.

retrain the next version of the model, closing the loop for a continuous learning strategy. Note that this differs from the classical approach to lifelong/continual learning, where it is assumed that access to previously available data is constrained or unavailable (Sodhani et al., 2022), as our new models have unrestrained access to all prior data.

Our choice of using GitHub Actions workflow stems from the ease of accessibility of previous spinal cord segmentation models in SCT (De Leener et al., 2017). When a new model is released on GitHub, it can be easily downloaded using the command `sct_deepseg_spinalcord -install -custom-url <release-url>` without having to install any model-specific packages. As a result, the GitHub Actions workflow is simply tasked with installing SCT and running the above-mentioned command for computing morphometrics across various models (accessible via the URL of their releases).

2.4. Validation protocol

2.4.1. Evaluation metrics

To evaluate the segmentation accuracy quantitatively, we report five metrics: (i) Jaccard index, (ii) Dice coefficient,

(iii) average surface distance (ASD), (iv) relative volume error (RVE), and (v) Hausdorff distance (HD) on the “frozen” test mentioned previously. The Jaccard Index measures the intersection over union of the prediction and the GT masks. It is known to be more sensitive to dissimilarity by false positives/negatives more harshly. While the Dice coefficient quantifies the overall agreement between input masks, Jaccard index quantifies what fraction of the prediction correctly overlaps with the GT. HD quantifies the worst-case boundary error (i.e., the largest of all the closest-point distances) unlike ASD, which quantifies the average of all closest-point distances between two surfaces. For clinically oriented assessment of the models, we also computed CSA averaged over C2–C3 vertebral levels of the cervical spinal cord on the predictions from the frozen test set defined previously to measure the morphometric variability for each model. The dataset characteristics of this test set including participant demographics, scanner manufacturers and field strengths are shown in Supplementary Table S2. These measurements are as follows:

1. **CSA:** The per-slice area (mm^2) of the predicted segmentation was computed across all slices and then averaged for each contrast.
2. **CSA STD:** For a given contrast, we computed the mean CSA over all slices averaged across the C2–C3 vertebral level. This was repeated for all contrasts for a given participant. Then, across all the participants, we computed the standard deviation (STD) of CSA across all contrasts to assess CSA variability.

The underlying assumption is that one participant should have similar spinal cord CSA across contrasts, with a lower CSA STD corresponding with a better model.

2.4.2. Evaluation of segmentations on various contrasts and pathologies

We compared the segmentations between our model’s current and previous versions to evaluate the quality of segmentations on challenging cases, including severely compressed spinal cords of DCM patients, and chronic hyperintense lesions of patients with SCI. We also evaluated our model’s ability to produce segmentations on MPRAGE T1map, resting-state axial gradient-echo echo-planar-imaging (GRE-EPI) on healthy participants and patients with cervical radiculopathy, whole-spine scans of healthy participants (Molinier et al., 2024) and scans acquired at 7T to highlight the model’s ability to generalize to various MRI contrasts, fields-of-view, scanner strengths, and pathologies unseen during training.

Table 1. Characteristics of the test set of spinal cord pathologies used to evaluate segmentation models.

Variables	MS	SCI	DCM
Number of participants	36	60	30
Sequences	T2*w	T2w	T2w
Number of MRI scans	36	60	39
Sex (Male/Female)	16 / 20	46 / 13 ^a	14 / 9 ^b
Age (y) (mean ± standard deviation)	N/A	53.0 ± 15.7	59.5 ± 11.5
Age range (y)	N/A	15 – 81	37 – 81
MRI manufacturers	Philips (<i>n</i> = 36)	GE (<i>n</i> = 37), Siemens (<i>n</i> = 21), N/A (<i>n</i> = 2)	Siemens (<i>n</i> = 36), N/A (<i>n</i> = 3)
MRI field strength	3T (<i>n</i> = 36)	1.5T (<i>n</i> = 49), 3T (<i>n</i> = 9), N/A (<i>n</i> = 2)	3T (<i>n</i> = 36), N/A (<i>n</i> = 3)
MRI sequence parameters	in-plane: [0.29 – 0.47] × [0.29 – 0.47] mm ² slice thickness: 2.5 – 5.0 mm	in-plane: [0.35 – 0.8] × [0.35 – 0.8] mm ² slice thickness: 2.75 – 6.0 mm	in-plane: [0.28 – 0.8] × [0.28 – 0.8] mm ² slice thickness: 0.8 – 3.6 mm

Note the heterogeneity in scanner manufacturers, field strengths, and image resolutions. MS = multiple sclerosis, SCI = spinal cord injury, DCM = degenerative cervical myelopathy.

^aAge/sex not reported for 1 SCI participant.

^bAge/sex not reported for 7 DCM participants.

We quantitatively compared the proposed model (`contrast_agnostic_v3.0`) with its predecessor (`contrast_agnostic_v2.0`) and existing open-source pathology-specific models `sct_deepseg_sc` (Gros et al., 2019) (for MS) and `SCIsegV2` (Karthik et al., 2024; Naga Karthik, Valošek, et al., 2025) (for SCI and DCM) using the Jaccard Index, Dice Coefficient, Relative Volume Error (RVE), Surface Distance, and Hausdorff Distance from the ANIMA toolbox (Commowick et al., 2018). These models were evaluated on a held-out test set strictly unseen during training following identical preprocessing. The dataset characteristics of pathology-specific datasets are given in Table 1.

2.4.3. Quantitative evaluation of morphometric drift

We applied the proposed lifelong learning framework and quantified the drift in the morphometric variability in terms of the STD of CSA across six contrasts (T2w, T1w, T2*w, MT-on, GRE-T1w, and DWI). Specifically, once released, we let the GitHub Actions workflow run the morphometric analysis and compare our proposed model against two existing spinal cord segmentation methods; `sct_deepseg_sc` (Gros et al., 2019), and `contrast_agnostic_v2.0` (Bédard et al., 2025).

2.4.4. Ablation study with recursively generated GT spinal cord masks

As described in Section 2.2, the spinal cord masks used as GT during training are gathered from multiple sites, containing a combination of manually annotated masks, masks obtained from automatic pathology-specific models. As a

result, the differences in delineating the spinal cord–CSF boundary might vary across individual expert raters and the automatic methods due to partial volume effects, hindering model performance. To eliminate this potential noise in the distribution of GT masks gathered from multiple sites, we performed an ablation study where the proposed model was used to produce new GT masks for the entire training set. In practice, this was achieved by running the inference on the entire training dataset and using the automatically generated predictions as the new GT masks for training the subsequent model without any manual corrections. As inter-rater biases are eliminated, the new set of GT masks represents a uniform distribution of GT labels.

2.4.5. Updating the normative database of spinal cord morphometrics

The database of healthy adult morphometrics proposed by Valošek et al. (2024) included morphometrics measures computed from 203 healthy individuals from the open-access Spine Generic Multi-Subject dataset (Cohen-Adad et al., 2021b). These morphometric measures were obtained from segmentations generated with `sct_deepseg_sc` (Gros et al., 2019), with manual corrections for over-/under-segmentation errors. As outlined in Section 1, morphometric measures are dependent on the segmentation method used. Therefore, we evaluated the following strategy of monitoring and updating the normative database:

1. Generate new segmentations using the proposed `contrast_agnostic_v3.0` model on the T2w

scans from 203 healthy participants in the normative database (Valošek et al., 2024).

2. Perform a manual quality control of the spinal cord segmentation masks.
3. Compute six morphometric measures (CSA, anteroposterior diameter, transverse diameter, compression ratio, eccentricity, and solidity) from the segmentation masks (Valošek et al., 2024).
4. Compute a scaling factor between the morphometric measures derived from different segmentation models, allowing for comparison of morphometric measures across segmentation models.

$$\text{Scaling Factor} = \frac{\text{metric}_{\text{contrast-agnostic_v3.0}}}{\text{metric}_{\text{sct_deepseg_sc}}}$$

2.4.6. Statistical analysis

Statistical analysis was performed using the SciPy (Virtanen et al., 2020), version 1.9.1, and scikit-posthocs (Terpilowski, 2019), version 0.10.0. Data normality was tested using the D'Agostino and Pearson's normality test. Between-group comparisons of CSA variability between the proposed and previous models were performed using the non-parametric Friedman test (following significant results from the normality test). Post hoc tests for pairwise between-group comparisons were performed using the Nemenyi test (following significant results from the Friedman test). To compare the significance of CSA variability between `contrast_agnostic_v3.0` and the model trained on recursively generated GT masks, we used the non-parametric pairwise Wilcoxon signed-rank test. Unless specified otherwise, $p < 0.001$ was considered to indicate a statistically significant difference.

3. RESULTS

3.1. Evaluation on various contrasts and pathologies

3.1.1. Qualitative comparison of segmentations

Figure 4 qualitatively compares the segmentations of `contrast_agnostic_v3.0` (current version), `contrast_agnostic_v2.0` (previous version), and `sct_deepseg_sc` on healthy and pathological scans. While all three models were trained on T1w, T2w, and T2*w contrasts, `contrast_agnostic_v2.0` was trained on healthy participant data only and `sct_deepseg_sc` was trained on a multisite dataset of MS patients. We observed a noticeable improvement in the segmentation of the heavily compressed spinal cord (with and without the presence of lesions) in DCM patients with our current model (`contrast_agnostic_v3.0`). Note that `contrast_agnostic_v1.0` is not a model but only a preliminary collection of scripts used to generate the soft ground truths (Bédard et al., 2025).

Figure 5 qualitatively shows the segmentation outputs of the model across a wide variety of contrasts and pathologies on both sagittal and axial orientations, including whole-spine scans. The model accurately segments the spinal cord under compression (DCM), in cases where the tubular structure of the cord is severely damaged (acute and chronic traumatic SCI) and in the presence of lesions (MS) and atrophy (ALS). All the images used for visualization belong to the test set gathered from different sites (as denoted by different participant IDs in the bottom left) and have never been encountered during training. Notably, in the case of whole-spine images, the model learned to segment

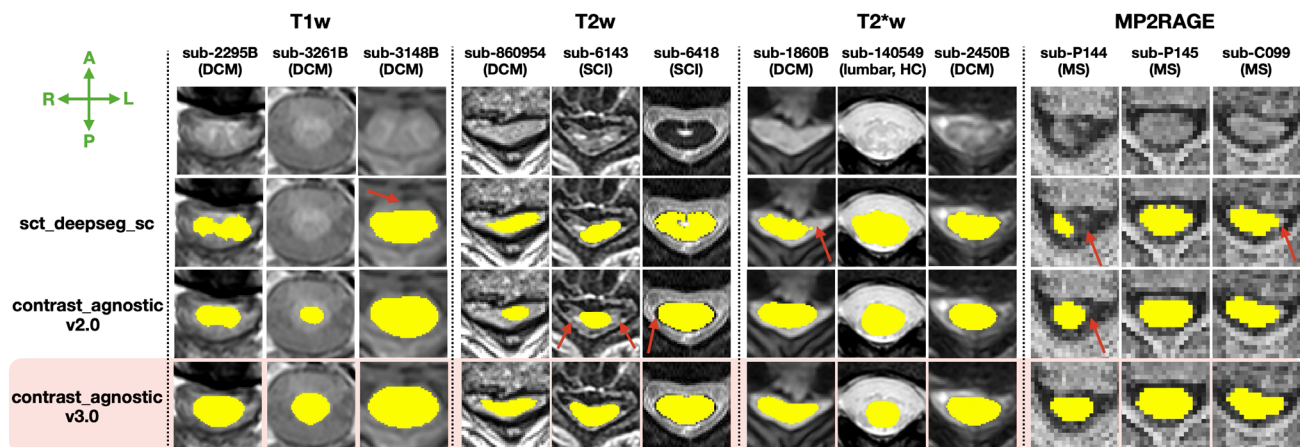


Fig. 4. Comparison of the automatic spinal cord segmentations between `contrast_agnostic_v3.0` (current version, highlighted), `contrast_agnostic_v2.0` (previous version), and `sct_deepseg_sc` on healthy controls (HC), DCM, SCI, and MS patients on the test set (unseen during training). Red arrows show the instances where the previous models fail, particularly under heavy compression (with/without lesions) in sub-860954, sub-6143, and sub-1860B.

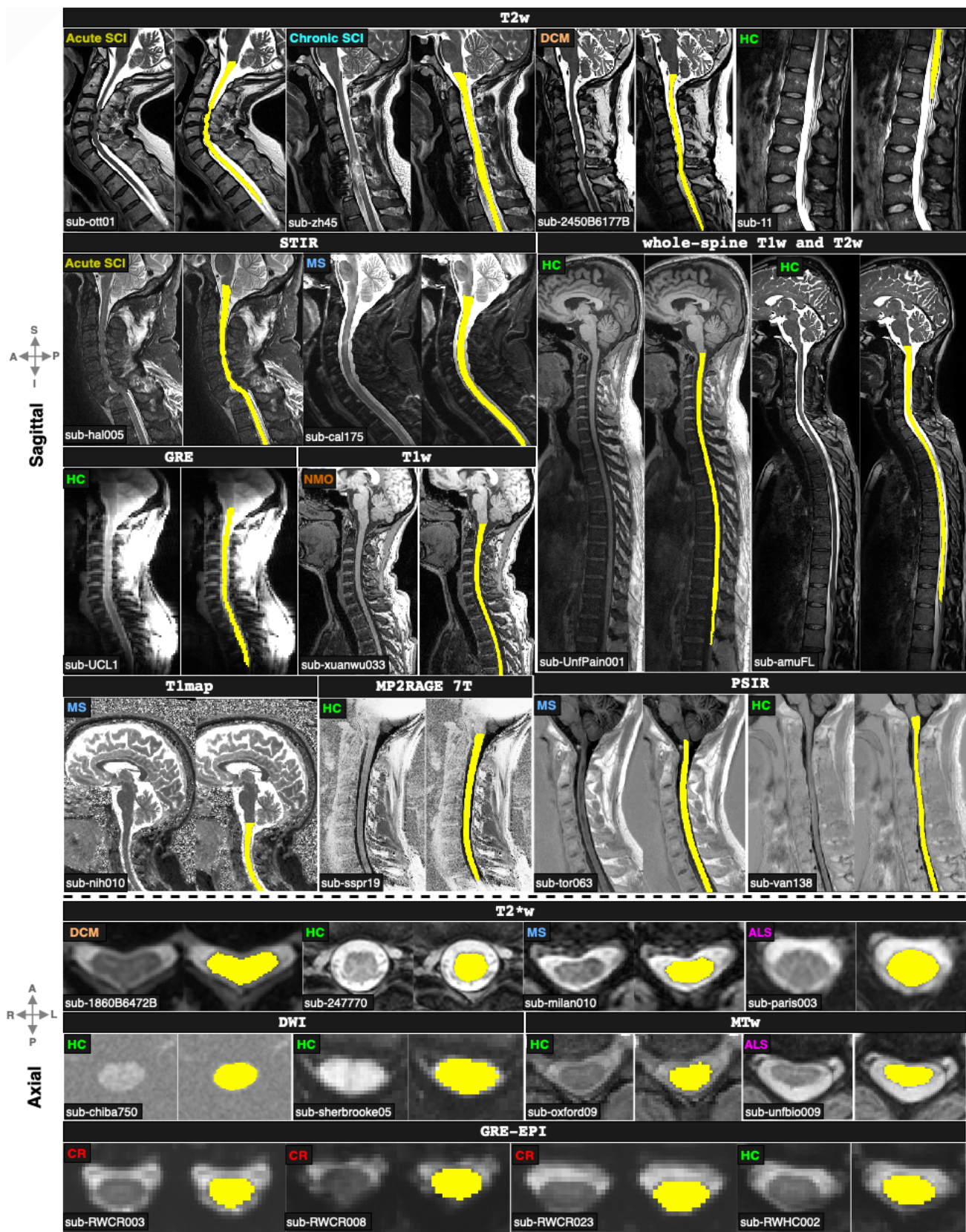


Fig. 5. Qualitative visualization of the proposed `contrast-agnostic_v3.0` model's segmentations across various contrasts and pathologies on test images from multiple sites. Our model accurately segments compressed spinal cords, severely damaged cords due to injury, and cords with the presence of lesions. Legend: SCI = spinal cord injury, DCM = degenerative cervical myelopathy, MS = multiple sclerosis, NMO = neuromyelitis optica, ALS = amyotrophic lateral sclerosis, CR = cervical radiculopathy, and HC = healthy control.

the entire spine despite only being trained independently on individual cervical, thoracic, and lumbar segments.

3.1.2. Quantitative evaluation on healthy controls and pathologies

Table 2 presents a quantitative comparison of the current (`contrast_agnostic_v3.0`) and previous (`contrast_agnostic_v2.0`) versions of the segmentation model in the lifelong training framework along with the existing pathology-specific models on test sets gathered from multiple sites containing healthy participant and pathological data. Starting with a comparison of the models on the frozen test set of healthy participants (Table 2A), we then present results for test sets containing T2w and T2*w images of the lumbar cord of healthy participants from two sites (Table 2B), T2*w images of MS patients from two sites (Table 2C), axial and sagittal T2w scans of DCM patients from two sites (Table 2D), and axial and sagittal T2w scans of traumatic SCI (acute, intermediate, and chronic phases) from six sites (Table 2E). In all comparisons, the proposed `contrast_agnostic_v3.0` model achieved similar or better performance compared

with the previous state-of-the-art or pathology-specific models with major improvement in the segmentation of the lumbar cord.

3.2. Quantitative evaluation of morphometric drift across model versions

3.2.1. Variability of CSA across contrasts

The figures below are automatically generated by the GitHub Actions workflow in the proposed lifelong training framework.

Figure 6 shows the CSA STD across six contrasts on the test set of healthy participants ($n = 49$; $n_{vol.} = 294$) of the spine-generic Multi-Subject database (Cohen-Adad et al., 2021b) between three methods: (i) `sct_deepseg_sc` (Gros et al., 2019), (ii) our previous version, `contrast_agnostic_v2.0` (Bédard et al., 2025), and the current version, `contrast_agnostic_v3.0`. The `contrast_agnostic_v3.0` model obtained relatively more stable segmentations with the lowest STD of CSA across contrasts compared with the other methods. Supplementary Figure S1 plots the variability in spinal cord CSA per each individual contrast. Similar to the analysis of CSA variability

Table 2. Quantitative comparison of spinal cord segmentations for previous segmentation methods on test sets containing: (A) healthy participants (cervical cord) ($n = 49$ participants; $n_{vol.} = 294$ images) averaged across all contrasts, (B) lumbar scans of healthy participants ($n = 9$), (C) patients with MS on T2*w contrast ($n = 36$ participants; $n_{vol.} = 36$ images), (D) DCM on axial and sagittal T2w scans ($n_{vol.} = 39$), and (E) SCI on axial and sagittal T2w scans ($n_{vol.} = 60$).

Methods	Jaccard (\uparrow)	Dice (\uparrow)	RVE (%)	ASD (\downarrow)	HD (\downarrow)
	Opt. value: 1	Opt. value: 1	Opt. value: 0	Opt. value: 0	Opt. value: 0
A) Healthy participants (cervical cord) ($n = 49$; 6 contrasts per participant; $n_{vol.} = 294$)					
<code>sct_deepseg_sc</code>	0.91 \pm 0.05	0.95 \pm 0.03	-0.18 \pm 8.95	0.04 \pm 0.27	3.18 \pm 6.16
<code>contrast_agnostic_v2.0</code>	0.91 \pm 0.03	0.95 \pm 0.02	-0.05 \pm 4.18	0.02 \pm 0.12	2.68 \pm 4.12
<code>contrast_agnostic_v3.0</code> (proposed)	0.92 \pm 0.03	0.96 \pm 0.02	-0.76 \pm 4.59	0.04 \pm 0.27	3.02 \pm 6.41
B) Healthy participants (lumbar cord) ($n = 9$; T2w and T2*w; $n_{vol.} = 9$)					
<code>sct_deepseg_sc</code>	0.67 \pm 0.13	0.78 \pm 0.11	57.56 \pm 53.28	0.34 \pm 0.58	61.87 \pm 101.46
<code>contrast_agnostic_v2.0</code>	0.60 \pm 0.24	0.71 \pm 0.26	32.47 \pm 32.94	7.14 \pm 18.58	57.04 \pm 84.45
<code>contrast_agnostic_v3.0</code> (proposed)	0.86 \pm 0.02	0.93 \pm 0.01	-3.61 \pm 4.78	0.015 \pm 0.02	7.48 \pm 7.42
C) Patients with MS ($n = 36$; T2*w contrast; $n_{vol.} = 36$)					
<code>sct_deepseg_sc</code>	0.89 \pm 0.03	0.94 \pm 0.02	-9.03 \pm 3.35	0.003 \pm 0.009	3.06 \pm 1.17
<code>contrast_agnostic_v2.0</code>	0.88 \pm 0.03	0.94 \pm 0.01	-10.12 \pm 2.89	0.009 \pm 0.016	3.43 \pm 1.61
<code>contrast_agnostic_v3.0</code> (proposed)	0.92 \pm 0.03	0.96 \pm 0.01	-5.34 \pm 2.89	0.005 \pm 0.014	2.96 \pm 2.39
D) Patients with DCM ($n = 32$; T2w contrast; $n_{vol.} = 39$)					
<code>SClsegV2</code>	0.94 \pm 0.02	0.97 \pm 0.01	-2.34 \pm 1.79	0.001 \pm 0.001	3.13 \pm 1.67
<code>contrast_agnostic_v2.0</code>	0.84 \pm 0.04	0.91 \pm 0.02	-11.91 \pm 4.16	0.01 \pm 0.04	6.23 \pm 4.83
<code>contrast_agnostic_v3.0</code> (proposed)	0.93 \pm 0.02	0.96 \pm 0.01	-2.51 \pm 2.25	0.001 \pm 0.001	3.14 \pm 1.62
E) Patients with SCI ($n = 60$; T2w contrast; $n_{vol.} = 60$)					
<code>SClsegV2</code>	0.86 \pm 0.06	0.93 \pm 0.04	5.22 \pm 7.63	0.01 \pm 0.01	12.83 \pm 23.51
<code>sct_deepseg_sc</code>	0.73 \pm 0.23	0.82 \pm 0.23	-13.68 \pm 24.1	7.61 \pm 31.87	38.39 \pm 77.35
<code>contrast_agnostic_v2.0</code>	0.60 \pm 0.17	0.74 \pm 0.17	-28.81 \pm 20.49	1.38 \pm 4.56	64.27 \pm 85.62
<code>contrast_agnostic_v3.0</code> (proposed)	0.87 \pm 0.08	0.93 \pm 0.06	1.75 \pm 14.63	0.01 \pm 0.04	10.3 \pm 22.6

RVE stands for relative volume error, ASD stands for average surface distance, and HD stands for Hausdorff distance. **Bold** values represent the best-performing model for that metric.

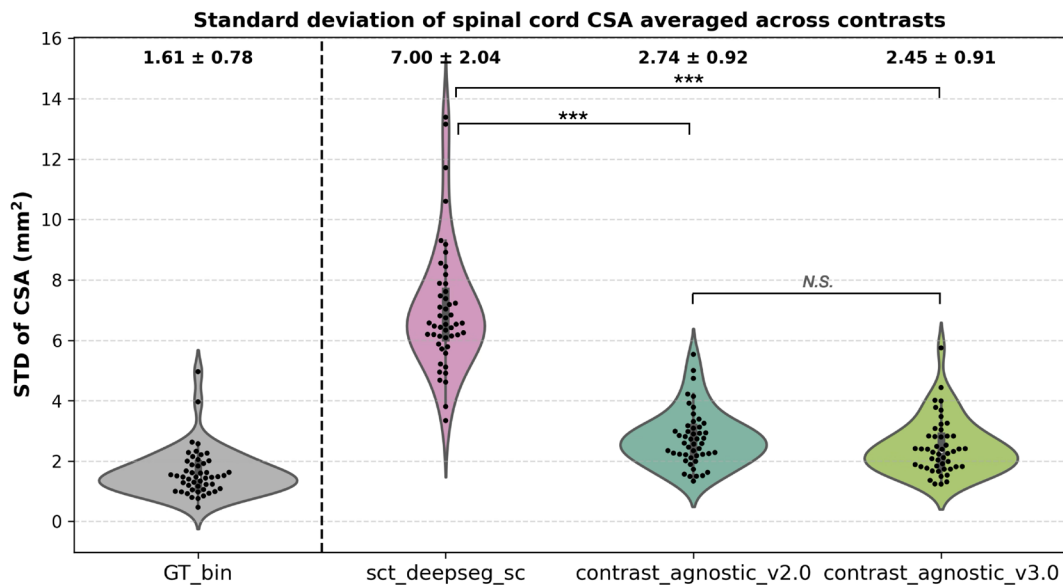


Fig. 6. CSA variability measured in terms of the standard deviation across 6 contrasts on a test set of healthy participants ($n = 49$). The lower the CSA STD across contrasts, the better. Our proposed model achieved the lowest STD averaged across six contrasts (i.e., each point shows the mean of six contrasts for a given participant) showing more stability in segmentations across contrasts. $***p < 0.001$ (non-parametric Friedman test followed by post hoc Nemenyi test for pairwise comparisons). “N.S.”: non-significant. Note that there is no statistically significant difference between `contrast_agnostic_v2.0` and `contrast_agnostic_v3.0`, implying that CSA variability achieved by the proposed model is at least as good as its previous version despite the addition of a diverse set of contrasts and pathologies.

across contrasts, we also plot the variability in CSA across 3 vendors (GE, Siemens, and Philips) on a test set containing scans of a healthy participant acquired from 15 sites in Supplementary Figure S2.

In Figure 7, we plot the level of agreement between the CSA estimated by the models on the commonly used T1w and T2w contrasts on the same test set described above. In addition to segmenting a wide range of contrasts and pathologies as shown in the previous figures, the `contrast-agnostic_v3.0` model achieves a similar alignment between T1w and T2w contrasts as our previous model trained only on a healthy participant database. Table 2A compares the models’ performances using quantitative metrics.

3.2.2. CSA variability with recursively generated GT masks

Since the GT masks for each contrast and pathology in the training set are a mixture of manual segmentations from different raters and automatic segmentations from different models, the collection of GT masks can be seen as a noisy distribution of segmentations with high variability at the spinal cord–CSF boundary. Figure 8 shows the results of our ablation study where all the GT masks were re-generated with `contrast-agnostic_v3.0`, and a new model was trained on the resulting collection. Recall that no manual corrections (or QC) were per-

formed to maintain a uniform distribution of the regenerated GT masks. We used the same test set of healthy participants ($n = 49$, 6 contrasts) from the spine-generic multi-subject database and compared two models: (i) the proposed model with the original (noisy) distribution of GT masks (shown with the green violin plot), and (ii) the proposed model, but trained on the new (uniform) distribution of the GT masks (shown with the blue violin plot). We observed that the model trained on the recursively generated GT masks showed a slightly higher STD across contrasts compared with the model trained on the original GT masks. In Supplementary Figure S3, we also plot the variability in CSA per contrast between the two methods, demonstrating how the model trained on recursively generated GT masks underestimated the CSA on all contrasts.

3.2.3. Normative database results

Figure 9A shows the plots for 6 different morphometric measures computed on 203 healthy participants using 2 versions of segmentation masks: (i) the segmentations from `sct_deepseg_sc` with manual corrections (pink) used in Valošek et al. (2024) and (ii) the segmentations from the proposed `contrast-agnostic_v3.0` model (green, no manual correction). Given the difference in the segmentations at the cord–CSF boundary, we present the *scaling factor* between the morphometric measures

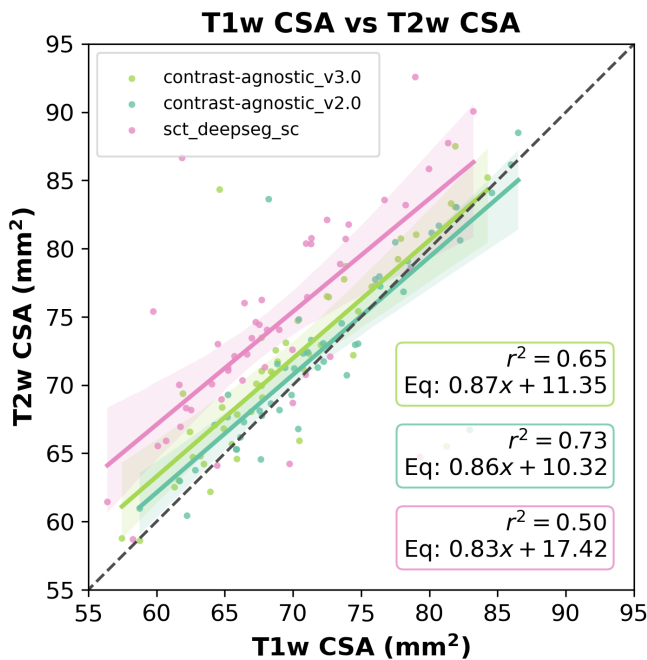


Fig. 7. Level of agreement between T1w and T2w CSA at C2–C3 for `contrast_agnostic_v3.0`, `contrast_agnostic_v2.0`, and `sct_deepseg_sc`. Each point represents one test participant ($n = 49$). The dashed diagonal line in black represents perfect agreement between the CSA of T1w and T2w contrasts. Points in the upper left triangle of the diagonal line represent overestimation of T2w CSA. Note the `sct_deepseg_sc` heavily overestimates T2w CSA for majority of test participants, whereas contrast agnostic models v2.0 and v3.0 lower the CSA variability.

computed with the two methods in Figure 9B. We observed that the scaling factor is nearly constant among slices across the given vertebral levels. For the benefit of future studies using the normative database of spinal cord morphometrics, they have been made open source.⁴

4. DISCUSSION

In this study, we presented an automatic model for the robust segmentation of the spinal cord across different MRI contrasts and pathologies. Our model was developed using heterogeneous data gathered from 75 clinical sites and hospitals worldwide, acquired with different resolutions, orientations, field strengths, and scanner manufacturers. We have shown that our proposed model provides reliable spinal cord segmentation on MRI scans across different pathologies including spinal cord compression (asymptomatic compression and DCM), atrophy (ALS), severely injured spinal cords in traumatic SCI, and spinal

cords containing intramedullary lesions (SCI and MS). To facilitate the continual development of segmentation models over time, we presented a *lifelong learning* scenario to automatically monitor the drift in morphometric variability across various model versions and enable periodic retraining by adding new contrasts and pathologies. As a real-world application of the lifelong learning framework, we applied the most recent version of our spinal cord segmentation model to update the morphometric measures of a normative database of healthy adults.

4.1. Data curation

Data gathered from multiple sites tend to be noisy in many respects, due in part to various imaging artifacts, metallic hardware, and environmental noise. While noisy GT masks are inevitable due to inter-rater variability, they could potentially be useful for training robust segmentation models (Shi & Wu, 2021; Yao et al., 2023). However, noise in training data tends to disrupt model training by making the models unintentionally focus on such outliers (Rahman et al., 2022; Taha & Hanbury, 2015), resulting in poor overall segmentation and inaccurate evaluation of the models' performance. In our proposed lifelong training scenario, it was critical to ensure the quality of the input data at each step of model development over time, as our segmentation models were trained from scratch on all the previous and new data. To account for this, we labeled each new dataset containing new contrasts or pathologies with existing automatic models (Bédard et al., 2025; Gros et al., 2019; Naga Karthik, Valošek, et al., 2025) and used `sct_qc` (SCT's visual QC tool) to quickly identify cases with failed segmentations requiring manual corrections and flagged images with strong artifacts for exclusion. These QC reports provide compressed snapshot of the datasets, which is useful for sharing with the clinical sites (Jwa et al., 2025).

Inter-rater variability was not quantitatively assessed in this study. Most manual segmentations were done by single raters at their respective sites with no traveling/overlapping participants/raters across sites. However, evidence from the literature shows that automated methods perform on par with the inter-rater discrepancy. De Leener et al. (2014) found PropSeg predictions and individual manual segmentations from two raters had similar Dice scores, comparable with inter-rater variability. In a clinical study, Connor et al. (2025) found 90.8% agreement between manual and automated measures of lesion damage following SCI on T2-w scans. In a similar set-up, Veiga-Canuto et al. (2022) compared manual and nnUNet-based automatic segmentations of neuroblastic tumors in MRI. They found the automatic method achieved similar inter-rater vari-

⁴<https://github.com/spinalcordtoolbox/PAM50-normalized-metrics/releases/tag/r20250321>

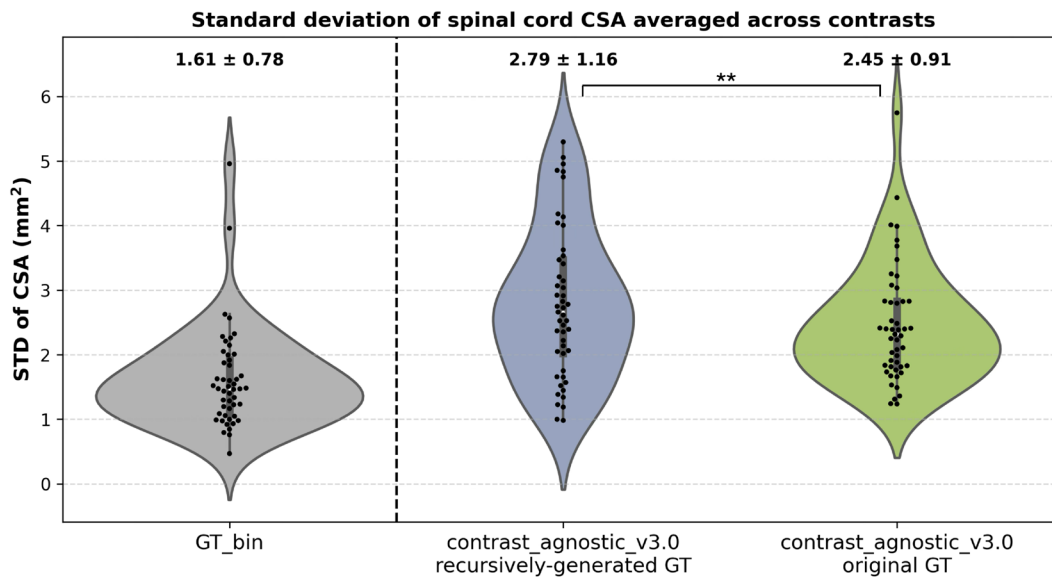


Fig. 8. Standard deviation of the CSA across six contrasts for models trained on (i) recursively generated GT masks (blue) and (ii) original GT masks (green). Each point shows the mean of six contrasts for a given participant. The lower the CSA STD across contrasts, the better. Note that the model trained on original GT masks (combination of manual and automatic segmentations) tends to produce stable segmentations resulting in a lower STD across contrasts than the uniform distribution of recursively generated GT masks. $**p < 0.01$ (two-sided non-parametric Wilcoxon signed-rank test).

ability to manual raters, with a median false-negative ratio of 0.96. Given the fact that the spinal cord is a large, easily detectable object compared with smaller objects such as lesions or tumors, we anticipate high agreement between the proposed `contrast_agnostic_v3.0` and inter-rater variability, consistent with prior research.

4.2. Lifelong learning segmentation of the spinal cord

4.2.1. Robustness across contrasts and pathologies

Gathering datasets containing new contrasts and pathologies over time, and training a model on this aggregated dataset resulted in robust segmentation of the spinal cord on a wide range of contrasts and pathologies. As seen in [Figures 4 and 5](#), the `contrast_agnostic_v3.0` model performed comparatively well when measured against the performance of previous models when applied to unseen images, benefitting from the lifelong learning strategy of updating the training database with new contrasts and pathologies. This was particularly notable for samples which exhibited severe compression, in the presence of both hyper-/hypo-intense lesions (MS and its phenotypes and different SCI phases), on lumbar spine, and unusual scanner strengths (7T MP2RAGE). Our model also performed well by generalizing to MRI contrasts not included in the training set (e.g., MPRAGE T1map, GRE-EPI, and Fieldmap images). Inter-

estingly, the model was also capable of accurate whole-spine segmentation, despite only being trained on “chunks” of individual spinal regions. This echoes the findings of our recent study, which found that segmentation models do not benefit from additional context when trained on scans covering the entire spinal cord ([Naga Karthik, McGinnis, et al., 2025](#)).

The competitive performance of the proposed model compared with existing pathology-specific models ([Table 2](#)) highlights the advantage of continually developing segmentation models over time as it reduces the cost of maintaining multiple models while ensuring that single class of models can be trained to be contrast- and pathology-agnostic over time.

4.2.2. Automatic monitoring of morphometric drift

Continuous monitoring of deployed models in production is a standard practice in MLOps pipelines, achieved through software technologies such as Docker, GitHub Actions, Kubernetes, and Git LFS ([Kandpal et al., 2023](#); [Spjuth et al., 2021](#); [Tabassam, 2023](#)). In a continuous learning system, monitoring deployed models is critical to ensure that the performance of the models on downstream tasks does not significantly degrade throughout their evolution ([Agirre et al., 2021](#)). Performance drifts could be caused by shifts in the input data distribution, typically manifesting in the form of changes in the participant demographics (e.g., adult population to pediatric population) and acquisition parameters (e.g., 3T data to 7T data)

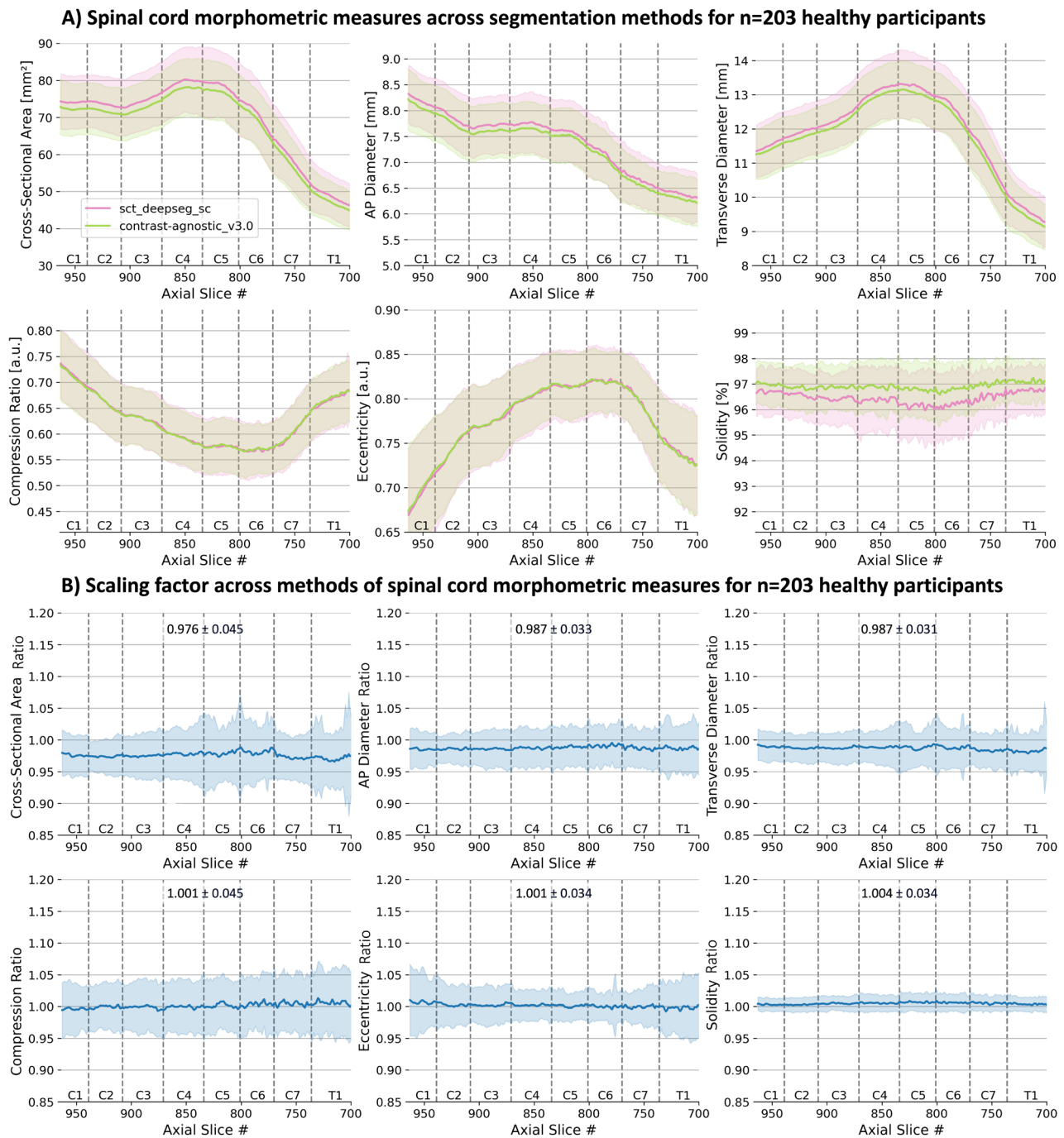


Fig. 9. (A) Morphometric measures computed on $n = 203$ healthy participants from the Spine Generic Dataset (Cohen-Adad et al., 2021b) for 6 morphometric measures using 2 different segmentation methods: `sct_deepseg_sc` with manual correction (green) and `contrast-agnostic_v3.0` (pink) with (B) scaling factor between the methods (means \pm std). Metrics are shown in the PAM50 space. Note that despite the diversity of the training dataset containing various contrasts and pathologies, the proposed v3.0 model achieved nearly constant scaling factors for all morphometric measures compared with the v2.0 model which was trained only on a healthy participant database.

(González et al., 2024). Therefore, monitoring morphometric drift between various model versions is crucial, as downstream tasks which rely on quantifying changes in the spinal cord morphometry are strongly tied to the accuracy of the segmentation (Joo et al., 2025; Valošek et al., 2024). In this regard, our proposed automatic workflow for

monitoring morphometric drift provides a quick feedback loop with two possible outcomes: (i) the magnitude of drift in the CSA variability with the new model is high, thus requiring re-evaluation of the data curation and/or model training steps to bring the drift within an acceptable range, or (ii) the magnitude of CSA drift is within an acceptable

range of the previous “stable” version, making it the new state-of-the-art for annotating (new) unlabeled data to train subsequent models. In the first outcome, we note that the CSA drift threshold was not defined a priori and that model retraining criteria were judged using CSA STD plots (Fig. 6) and qualitative evaluation of model’s predictions at the cord boundaries with QC reports. As the test set required to be publicly accessible for the GitHub workflow and only contain the images of healthy participants from six contrasts to ensure fair comparison with previous model, we observed that additional QC in challenging pathological cases such as the ones with lesions and/or heavily compressed cords was critical to ensure that the model learned to segment the spinal cord boundaries appropriately.

Another important note is that the morphometric drift was primarily assessed by measuring the CSA across contrasts. While antero-posterior (AP) and right-left (RL) diameters could also capture changes related to boundary positioning, they would essentially reflect similar underlying information as CSA. Other shape-based metrics such as eccentricity or compression ratio measure the overall shape of the cord (i.e., its ellipticity) or ratios of AP/RL diameters but are not suitable for measuring systematic shifts in cord boundaries.

Also, note that the proposed lifelong learning framework using GitHub Actions is not specific to spinal cord segmentation but can be reused for any other segmentation task involving the development of multiple models over time.

4.2.3. Training on recursively generated labels

Any form of human intervention is undesirable in a post-deployment lifelong learning scenario making it prone to errors. However, existing models are unable to automatically utilize incoming data as it arrives (Agirre et al., 2021; González et al., 2024; Prapas et al., 2021), necessitating periodic checks to prevent degradation of model performance. While our proposed continuous training strategy automatically monitors the drift in morphometric variability after training, one could also automate the re-training process, thus making the continuous learning loop *fully automatic*. Currently, when new data arrive, we rely on the combination of automatic annotation using the latest version of the model and performing visual QC, identifying cases with failed/incorrect segmentations for manual corrections. What if we forego this data curation step involving manual intervention?

In our attempt to evaluate the potential of such an approach (Fig. 8; Supplementary Fig. S3), we observed that the model underestimated the average CSA on a healthy subset of participants for each of the six contrasts

and resulted in a significantly higher CSA STD across contrasts ($p < 0.01$), when compared with the performance of the model trained on the original GT masks obtained from a combination of automatic and manual segmentations. Recent research in the context of text generation and image synthesis (Shumailov et al., 2024) has shown that multiple iterations of training on recursively generated data tend to make the model *catastrophically forget* (Sodhani et al., 2022) the underlying true data distribution, leading to model collapse, something which we did not observe. Given the inconsistencies in manual/automatic segmentations at the cord-CSF boundary owing to varying partial volume effects with images of different contrasts and resolutions, we hypothesize that training on such noisy labels acted as an inherent regularizer, making the model more robust across contrasts. However, training on uniform distribution of model-generated segmentations where the inconsistencies have been smoothed out, the model tends to under-segment the spinal cord, something which would need to be kept in mind for analyses based on models trained this way.

Successive training with recursively generated GT masks can potentially propagate errors in the training data distribution. To quantitatively assess bias introduced by recursive use of automatic segmentations, an initial baseline ($X\%$) is established by evaluating the original model (M_0) on a manually annotated test set (T_0). A subsequent model (M_1) is then trained using masks automatically generated by M_0 and evaluated on (i) the auto-generated masks ($Y\%$) and (ii) the original manual set (T_0) ($Z\%$). The discrepancy ($Y-Z\%$) quantifies the degree of bias, with larger values indicating that M_1 has overfit to systematic errors in the recursive ground truth. This experiment is difficult to scale, as fully manual annotations remain time consuming and costly to produce.

4.2.4. Binary versus soft masks

While training directly on soft masks still achieves the lowest morphometric variability across contrasts (Bédard et al., 2025), the registration step (which requires mutual co-registration of all contrasts) requires more than one contrast per participant, becoming a bottleneck in developing segmentation models, as well as further manual intervention in correcting registration outputs across both healthy and pathological data. Furthermore, training on *soft* masks requires converting existing datasets with binary GT masks to soft masks within an appropriate contrast-dependent threshold. Given the lifelong learning framework for developing segmentation models, the *softness* of the masks from one model cannot be accurately quantified to match the softness for the next model, owing to partial volume effects and differences in the training data distribution, sub-

tly biasing the ground truth with subsequent newer versions of the model. On the contrary, training on binarized GT masks (thresholded at 0.5) presents a simple and scalable solution, reducing the impact of model-specific biases as most models tend to be uncertain at the boundaries of the segmentation masks (Lemay et al., 2022). While training on binary masks is scalable in a lifelong learning framework, it could potentially be limiting in cases where the CSA is small at the tip of the spinal cord. In these regions, soft masks can better represent the partial volume compared with binary masks. We refer the reader to Bédard et al. (2025) for detailed qualitative and quantitative comparisons on the effects of training with soft and binary masks and the usage of pixel-wise classification and regression-based loss functions toward the estimation of CSA.

4.3. Application on normative database of morphometrics

Keeping an updated normative morphometrics database is crucial to maintaining lifelong models (Valošek et al., 2024), as it allows users to relate their measurements obtained using the latest segmentation method up-to-date. Additionally, when adding new individuals to the normative database, one should re-segment all images within it using the latest segmentation method to ensure the database follows the state of the model. Maintaining and updating such a dataset requires coordination across the segmentation model, the SCT software, and the Spine Generic dataset, a process not currently implemented, but can be accomplished using GitHub Actions. The scaling factors identified using our framework also ensures backward compatibility with previous segmentation methods included in SCT (i.e., `sct_deepseg_sc`), allowing researchers to compare morphometric measures derived from different segmentation models. We encourage users to update to `contrast_agnostic_v3.0`, however, as it significantly improves the spinal cord segmentation robustness in previously difficult pathologies, such as cord compression and spinal cord injury.

4.4. Limitations

A major limitation of this study is that our strategy for monitoring and evaluating morphometric drift across various model versions depends on a fixed set of contrasts ($n = 6$) in a frozen test set of healthy participants. While newer models may generalize well to other pathologies and contrasts, their true performance could be limited by the evaluation of the CSA on only six contrasts. Future work could add better methods for evaluating morphometric drift (e.g., by computing other commonly used

spinal cord morphometrics) on data from both healthy participants and from participants with spinal cord pathologies. With the rise of open-source challenges targeting specific spinal cord pathologies,^{5,6} our GitHub Actions-based workflows could be adapted to include evaluations not only of healthy participants but of participants with pathologies as well.

Stagnation of the training data distribution when developing models over time is another issue. With subsequent models being trained on new data (potentially from different populations—pediatric, adult, and geriatric), the data distribution used for the earliest model might no longer be representative of the current distribution. In such cases, comparing histogram-based distribution shifts using KL divergence, or detecting drifts in the feature space by extracting radiomic features (van Griethuysen et al., 2017), could ensure the continued relevance of the training and test sets for evaluating future models. If the drift between data distribution is large, keeping only a subset of the old data when training new models is recommended.

5. CONCLUSION

This study introduces an automatic tool for the robust segmentation of the spinal cord across various MRI contrasts and spinal pathologies. The model was trained on diverse datasets collected from 75 clinical sites and hospitals worldwide, with heterogeneous image resolutions, orientations, field strengths, and scanner manufacturers. Our results demonstrate that the model effectively segments spinal cord scans from healthy participants, as well as from those with compressions, atrophy, intramedullary lesions, and SCI. To support the continuous improvement of segmentation models, we propose a lifelong learning framework which automatically monitors the drifts in morphometric variability across model versions. The proposed framework facilitates periodic retraining by incorporating new contrasts and pathologies and provides a quick feedback loop for developing future segmentation models. As a real-world application of this framework, we employed the proposed spinal cord segmentation model to update morphometric measurements in a normative database of healthy adults. Our results showed that the scaling factor required to update the database of morphometric measures is nearly constant among slices across the given vertebral levels, showing minimum drift between the current and previous versions of the model trained within the lifelong learning framework.

⁵<https://portal.fli-iam.irisa.fr/ms-multi-spine/>

⁶<https://ivdm3seg.weebly.com>

DATA AND CODE AVAILABILITY

The datasets used in this study consist of a mix of publicly available (✓) and privately held datasets (×). The dataset identifiers used internally along with the accessibility are as follows: basel-mp2rage (×), canproco (×), data-multi-subject (×), dcm-brno (×), dcm-zurich-lesions-20231115 (×), dcm-zurich-lesions (×), dcm-zurich (×), lumbar-epfl (×), lumbar-vanderbilt (×), nih-ms-mp2rage (×), sci-colorado (×), sci-paris (×), sci-zurich (×), sct-testing-large (×), site_006_praxis (×), site_007_praxis (×). Publicly available datasets are cited using their corresponding publication wherever applicable. The code for this study is open source and can be accessed at <https://github.com/sct-pipeline/contrast-agnostic-softseg-spinalcord/releases/tag/v3.0> (which also contains the overview of dataset characteristics). For ease of accessibility, the spinal cord segmentation model is integrated into Spinal Cord Toolbox (v7.0 and above). The updated database of spinal cord morphometric measures can be accessed at <https://github.com/spinalcordtoolbox/PAM50-normalized-metrics/releases/tag/r20250321>.

AUTHOR CONTRIBUTIONS

E.N.K.: data curation, formal analysis, investigation, methodology, visualization, and writing (original draft, review, and editing). S.B.: data curation, formal analysis, investigation, methodology, visualization, and writing (original draft, review, and editing). J.V.: data curation, formal analysis, investigation, methodology, visualization, and writing (original draft, review, and editing). J.C.A.: conceptualization, data curation, funding acquisition, investigation, methodology, supervision, writing (review and editing). The rest of the authors: data curation and writing (review and editing).

FUNDING

E.N.K. is supported by the Fonds de Recherche du Québec Nature and Technologie (FRQNT) Doctoral Training Scholarship. S.B. is supported by the Natural Sciences and Engineering Research Council of Canada, NSERC, Canada Graduate Scholarships—Doctoral program. J.V. received funding from the European Union's Horizon Europe research and innovation program under the Marie Skłodowska–Curie grant agreement No 101107932. 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 TransMed-Tech), the Courtois NeuroMod project, the Quebec Bio-Imaging Network [5886, 35450], INSPIRED (Spinal Research, UK; Wings for Life, Austria; Craig H. Neilsen Foundation, USA), Mila—Tech Transfer Funding Program, the Digital Research Alliance of Canada (alliancecan.ca) for providing the supercomputing infrastructure, NIH R01NS078322 (CM). J.V., J.B., T.H., and P.K. received funding from Czech Health Research Council grants NV18-04-00159, NU22-04-00024, and LM2018129 Czech-BioImaging. D.V.D.V., I.R., and N.K. received funding from the Swiss National Science Foundation grant [205321_207493]. D.P. is funded by a Swiss National Science Foundation Postdoc Mobility Fellowship grant [P500PM_214211]. C.T. is funded by the Swiss National Science Foundation (Grant numbers: P500PM_206620 and P5R5PM_230625) and the National Multiple Sclerosis Society (Grant number: FG-2107-38022). U.H. and F.E. are supported by the Max Planck Society and the European Research Council [ERC StG 758974]. The Rick Hansen Spinal Cord Injury Registry is supported by funding from the Praxis Spinal Cord Institute through the Government of Canada and the Province of British Columbia. D.S.R. and G.N. are funded by the Intramural Research Program of the National Institute of Neurological Disorders and Stroke, NIH, USA.

ETHICS

All participants in their respective imaging sites signed a consent form as per the local IRB regulations. Data were then aggregated at the managing site, under Polytechnique Montréal's IRB (CER-2324-26-D).

DECLARATION OF COMPETING INTEREST

Conflicts of interest outside the present work: M. Filippi is editor-in-chief of the *Journal of Neurology*, associate editor of *Human Brain Mapping*, *Neurological Sciences*, and *Radiology*, received compensation for consulting services from Alexion, Almirall, Biogen, Merck, Novartis, Roche, Sanofi, speaking activities from Bayer, Biogen, Celgene, Chiesi Italia SpA, Eli Lilly, Genzyme, Janssen, Merck-Serono, Neopharmed Gentili, Novartis, Novo Nordisk, Roche, Sanofi, Takeda, and TEVA, participation in advisory boards for Alexion, Biogen, Bristol-Myers Squibb, Merck, Novartis, Roche, Sanofi, Sanofi-Aventis, Sanofi-Genzyme, Takeda, scientific direction of educational events for Biogen, Merck, Roche, Celgene, Bristol-Myers Squibb, Lilly, Novartis, Sanofi-Genzyme, he receives research support from Biogen Idec, Merck-Serono, Novartis, Roche, the Italian Ministry of Health,

the Italian Ministry of University and Research, and Fondazione Italiana Sclerosi Multipla.

M.A. Rocca received consulting fees from Biogen, Bristol Myers Squibb, Eli Lilly, Janssen, Roche; and speaker honoraria from AstraZeneca, Biogen, Bristol Myers Squibb, Bromatech, Celgene, Genzyme, Horizon Therapeutics Italy, Merck Serono SpA, Novartis, Roche, Sanofi and Teva. She receives research support from the MS Society of Canada, the Italian Ministry of Health, the Italian Ministry of University and Research, and Fondazione Italiana Sclerosi Multipla. She is associate editor for *Multiple Sclerosis and Related Disorders*; and associate co-editor for Europe and Africa for *Multiple Sclerosis Journal*.

E. Pravatà received honoraria for serving on scientific advisory boards from Bayer AG, unrelated to the present work. C. Mainero and C.A. Treaba received research support from Genentech. Daniel S. Reich has received research funding from Abata and Sanofi. The rest of the authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ACKNOWLEDGMENTS

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, and Kalum Ost for reviewing and editing this manuscript. We thank Drs. Thierry Albert, Bertrand Baussart, Caroline Huger, 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 Drs. Jean Pelletier and Bertrand Audoin, and Lauriane Pini, for data acquisition in Marseille. Finally, the authors would like to thank the RHSCIR participants and network, including all the participating local RHSCIR sites: Vancouver General Hospital, Foothills Hospital, Royal University Hospital, Toronto Western Hospital, St Michael's Hospital, Sunnybrook Health Sciences Centre, Hamilton General Hospital, The Ottawa Hospital Civic Campus, Hôpital de l'Enfant Jésus, Hôpital du Sacre Coeur de Montréal, QEII Health Sciences Centre and Saint John Regional Hospital. Data for RHSCIR were collected using the Global Research Platform <http://www.rhigrp.net/>. Neurology data for this study were cleaned using the Praxis ISNCSCI Algorithm <http://www.isncscialgorithm.com/>. We would like to acknowledge the Alberta Trauma Services Team and the Alberta Trauma Registry in providing data for the Rick Hansen Spinal Cord Injury Registry (Rotem-Kohavi et al., 2025).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available with the online version here: <https://doi.org/10.1162/IMAG.a.1105>.

REFERENCES

- Agirre, E., Jonsson, A., & Larcher, A. (2021). Framing lifelong learning as autonomous deployment: Tune once live forever. In *Lecture Notes in Electrical Engineering* (pp. 331–336). Springer. https://doi.org/10.1007/978-981-15-9323-9_29
- Alla, S., & Adari, S. K. (2020). *Beginning MLOps with MLFlow: Deploy models in AWS SageMaker, Google cloud, and Microsoft azure* (1st ed.). Apress. <https://doi.org/10.1007/978-1-4842-6549-9>
- Bautin, P., & Cohen-Adad, J. (2021). Minimum detectable spinal cord atrophy with automatic segmentation: Investigations using an open-access dataset of healthy participants. *NeuroImage. Clinical*, 32, 102849. <https://doi.org/10.1016/j.nicl.2021.102849>
- Bédard, S., & Cohen-Adad, J. (2022). Automatic measure and normalization of spinal cord cross-sectional area using the pontomedullary junction. *Frontiers in Neuroimaging*, 1, 1031253. <https://doi.org/10.3389/fnimg.2022.1031253>
- Bédard, S., Karthik, E. N., Tsagkas, C., Pravatà, E., Granziera, C., Smith, A., Weber, K. A., Li, & Cohen-Adad, J. (2025). Towards contrast-agnostic soft segmentation of the spinal cord. *Medical Image Analysis*, 101, 103473. <https://doi.org/10.1016/j.media.2025.103473>
- Bédard, S., Valošek, J., Seif, M., Curt, A., Schading, S., Pfender, N., Freund, P., Hupp, M., & Cohen-Adad, J. (2024). Normalizing spinal cord compression morphometric measures: Application in degenerative cervical myelopathy. *medRxiv*. <https://doi.org/10.1101/2024.03.13.24304177>
- Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv [cs.CV]*. <http://arxiv.org/abs/1706.05587>
- Chen, M., Carass, A., Oh, J., Nair, G., Pham, D. L., Reich, D. S., & Prince, J. L. (2013). Automatic magnetic resonance spinal cord segmentation with topology constraints for variable fields of view. *NeuroImage*, 83, 1051–1062. <https://doi.org/10.1016/j.neuroimage.2013.07.060>
- Cohen-Adad, J., Alonso-Ortiz, E., Abramovic, M., Arneitz, C., Atcheson, N., Barlow, L., Barry, R. L., Barth, M., Battiston, M., Büchel, C., Budde, M., Callot, V., Combes, A. J. E., De Leener, B., Descoteaux, M., de Sousa, P. L., Dostál, M., Doyon, J., Dvorak, A., ... Xu, J. (2021a). Generic acquisition protocol for quantitative MRI of the spinal cord. *Nature Protocols*, 16(10), 4611–4632. <https://doi.org/10.1038/s41596-021-00588-0>
- Cohen-Adad, J., Alonso-Ortiz, E., Abramovic, M., Arneitz, C., Atcheson, N., Barlow, L., Barry, R. L., Barth, M., Battiston, M., Büchel, C., Budde, M., Callot, V., Combes, A. J. E., De Leener, B., Descoteaux, M., de Sousa, P. L., Dostál, M., Doyon, J., Dvorak, A., ... Xu, J. (2021b). Open-access quantitative MRI data of the spinal cord and reproducibility across participants, sites and manufacturers. *Scientific Data*, 8(1), 219. <https://doi.org/10.1038/s41597-021-00941-8>
- Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S. C., Girard, P., Améli, R., Ferré, J.-C.,

- Kerbrat, A., Tourdias, T., Cervenansky, F., Glatard, T., Beaumont, J., Doyle, S., Forbes, F., Knight, J., Khademi, A., ... Barillot, C. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific Reports*, 8(1), 13650. <https://doi.org/10.1038/s41598-018-31911-7>
- Connor, J. R., Thornton, W. A., Weber, K. A., Pfyffer, D., Freund, P., Tefertiller, C., & Smith, A. C. (2025). Reliability of SCLseg automated measurement of midsagittal tissue bridges in spinal cord injuries using an external dataset. *Topics in Spinal Cord Injury Rehabilitation*, 31(2), 39–49. <https://doi.org/10.46292/sci25-00015>
- De Leener, B., Kadoury, S., & Cohen-Adad, J. (2014). Robust, accurate and fast automatic segmentation of the spinal cord. *NeuroImage*, 98, 528–536. <https://doi.org/10.1016/j.neuroimage.2014.04.051>
- De Leener, B., Lévy, S., Dupont, S. M., Fonov, V. S., Stikov, N., Louis Collins, D., Callot, V., & Cohen-Adad, J. (2017). SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data. *NeuroImage*, 145(Pt A), 24–43. <https://doi.org/10.1016/j.neuroimage.2016.10.009>
- Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., & Heng, P.-A. (2017). 3D deeply supervised network for automated segmentation of volumetric medical images. *Medical Image Analysis*, 41, 40–54. <https://doi.org/10.1016/j.media.2017.05.001>
- González, C., Fuchs, M., Santos, D. P. dos, Matthies, P., Trenz, M., Grüning, M., Chaudhari, A., Larson, D. B., Othman, A., Kim, M., Nensa, F., & Mukhopadhyay, A. (2024). Regulating radiology AI medical devices that evolve in their lifecycle. *arXiv [cs.CV]*. <http://arxiv.org/abs/2412.20498>
- Gros, C., De Leener, B., Badji, A., Maranzano, J., Eden, D., Dupont, S. M., Talbott, J., Zhuoquiong, R., Liu, Y., Granberg, T., Ouellette, R., Tachibana, Y., Hori, M., Kamiya, K., Chougar, L., Stawiarz, L., Hillert, J., Bannier, E., Kerbrat, A., ... Cohen-Adad, J. (2019). Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. *NeuroImage*, 184, 901–915. <https://doi.org/10.1016/j.neuroimage.2018.09.081>
- Horáková, M., Horák, T., Valošek, J., Rohan, T., Korit'áková, E., Dostál, M., Kočica, J., Skutil, T., Keřkovský, M., Kadaňka, Z., Jr, Bednařík, P., Svátková, A., Hlušík, P., & Bednařík, J. (2022). Semi-automated detection of cervical spinal cord compression with the Spinal Cord Toolbox. *Quantitative Imaging in Medicine and Surgery*, 12(4), 2261–2279. <https://doi.org/10.21037/qims-21-782>
- Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211. <https://doi.org/10.1038/s41592-020-01008-z>
- Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., & Jaeger, P. F. (2024). NnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. *arXiv [cs.CV]*. <https://doi.org/10.48550/ARXIV.2404.09556>
- Joo, B., Park, H. J., Park, M., Suh, S. H., & Ahn, S. J. (2025). Establishing normative values for entire spinal cord morphometrics in East Asian young adults. *Korean Journal of Radiology: Official Journal of the Korean Radiological Society*, 26(2), 146–155. <https://doi.org/10.3348/kjr.2024.0907>
- Jwa, A. S., Norgaard, M., & Poldrack, R. A. (2025). Can I have your data? Recommendations and practical tips for sharing neuroimaging data upon a direct personal request. *Imaging Neuroscience*, 3, imag_a_00508. https://doi.org/10.1162/imag_a_00508
- Kandpal, N., Lester, B., Muqeeth, M., Mascarenhas, A., Evans, M., Baskaran, V., Huang, T., Liu, H., & Raffel, C. (2023). Git-Theta: A Git extension for collaborative development of machine learning models. In *arXiv [cs.LG]*. <http://arxiv.org/abs/2306.04529>
- Karthik, E. N., Kerbrat, A., Labauge, P., Granberg, T., Talbott, J., Reich, D. S., Filippi, M., Bakshi, R., Callot, V., Chandar, S., & Cohen-Adad, J. (2022). Segmentation of Multiple Sclerosis lesions across hospitals: Learn continually or train from scratch? *arXiv [cs.CV]*. <http://arxiv.org/abs/2210.15091>
- Karthik, E. N., Valošek, J., Farnier, L., Pfyffer, D., Schading-Sassenhausen, S., Leuret, A., David, G., Smith, A. C., Weber, K. A., II, Seif, M., RHSCIR Network Imaging Group, Freund, P., & Cohen-Adad, J. (2024). SCLsegV2: A universal tool for segmentation of intramedullary lesions in spinal cord injury. *arXiv [cs.CV]*. <http://arxiv.org/abs/2407.17265>
- Kato, F., Yukawa, Y., Suda, K., Yamagata, M., & Ueta, T. (2012). Normal morphology, age-related changes and abnormal findings of the cervical spine. Part II: Magnetic resonance imaging of over 1,200 asymptomatic subjects. *European Spine Journal: Official Publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*, 21, 1499–1507. <https://link.springer.com/article/10.1007/s00586-012-2176-4>
- Labounek, R., Bondy, M. T., Paulson, A. L., Bédard, S., Abramovic, M., Alonso-Ortiz, E., Atcheson, N. T., Barlow, L. R., Barry, R. L., Barth, M., Battiston, M., Büchel, C., Budde, M. D., Callot, V., Combes, A., Leener, B. D., Descoteaux, M., Loureiro de Sousa, P., Dostál, M., ... Nestrašil, I. (2024). Body size interacts with the structure of the central nervous system: A multi-center in vivo neuroimaging study. *bioRxiv*. <https://doi.org/10.1101/2024.04.29.591421>
- Lemay, A., Gros, C., Karthik, E. N., & Cohen-Adad, J. (2022). Label fusion and training methods for reliable representation of inter-rater uncertainty. *arXiv [eess.IV]*. <http://arxiv.org/abs/2202.07550>
- Liu, B., & Mazumder, S. (2021). Lifelong and continual learning dialogue systems: Learning during conversation. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 35(17), 15058–15063. <https://doi.org/10.1609/aaai.v35i17.17768>
- Losseff, N. A., Webb, S. L., O'Riordan, J. I., Page, R., Wang, L., Barker, G. J., Tofts, P. S., McDonald, W. I., Miller, D. H., & Thompson, A. J. (1996). Spinal cord atrophy and disability in multiple sclerosis. A new reproducible and sensitive MRI method with potential to monitor disease progression. *Brain: A Journal of Neurology*, 119(Pt 3), 701–708. <https://doi.org/10.1093/brain/119.3.701>
- Lukas, C., Sombekke, M. H., Bellenberg, B., Hahn, H. K., Popescu, V., Bendfeldt, K., Radue, E. W., Gass, A., Borgwardt, S. J., Kappos, L., Naegelin, Y., Knol, D. L., Polman, C. H., Geurts, J. J. G., Barkhof, F., & Vrenken, H. (2013). Relevance of spinal cord abnormalities to clinical disability in multiple sclerosis: MR imaging findings in a large cohort of patients. *Radiology*, 269(2), 542–552. <https://doi.org/10.1148/radiol.13122566>
- Martin, A. R., De Leener, B., Cohen-Adad, J., Kalsi-Ryan, S., Cadotte, D. W., Wilson, J. R., Tetreault, L., Nouri, A., Crawley, A., Mikulis, D. J., Ginsberg, H., Massicotte, E. M., & Fehlings, M. G. (2018). Monitoring

- for myelopathic progression with multiparametric quantitative MRI. *PLoS One*, 13(4), e0195733. <https://doi.org/10.1371/journal.pone.0195733>
- Masse-Gignac, N., Flórez-Jiménez, S., Mac-Thiong, J.-M., & Duong, L. (2023). Attention-gated U-Net networks for simultaneous axial/sagittal planes segmentation of injured spinal cords. *Journal of Applied Clinical Medical Physics*, 24(10), e14123. <https://doi.org/10.1002/acm2.14123>
- Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *arXiv [cs.CV]*. <http://arxiv.org/abs/1606.04797>
- Molinier, N., Bédard, S., Boudreau, M., Cohen-Adad, J., Callot, V., Alonso-Ortiz, E., Pageot, C., & Laines-Medina, N. (2024). "whole-spine" [Dataset]. OpenNeuro. <https://doi.org/10.18112/openneuro.ds005616.v1.0.1>
- Naga Karthik, E., McGinnis, J., Wurm, R., Ruehling, S., Graf, R., Valosek, J., Benveniste, P.-L., Lauerer, M., Talbott, J., Bakshi, R., Tauhid, S., Shepherd, T., Berthele, A., Zimmer, C., Hemmer, B., Rueckert, D., Wiestler, B., Kirschke, J. S., Cohen-Adad, J., & Mühlau, M. (2025). Automatic segmentation of spinal cord lesions in MS: A robust tool for axial T2-weighted MRI scans. *medRxiv*. <https://doi.org/10.1101/2025.01.22.25320959>
- Naga Karthik, E., Valošek, J., Smith, A. C., Pfyffer, D., Schading-Sassenhausen, S., Farner, L., Weber, K. A., 2nd, Freund, P., & Cohen-Adad, J. (2025). SClseg: Automatic segmentation of intramedullary lesions in spinal cord injury on T2-weighted MRI scans. *Radiology. Artificial Intelligence*, 7(1), e240005. <https://doi.org/10.1148/ryai.240005>
- Nozawa, K., Maki, S., Furuya, T., Okimatsu, S., Inoue, T., Yunde, A., Miura, M., Shiratani, Y., Shiga, Y., Inage, K., Eguchi, Y., Ohtori, S., & Orita, S. (2023). Magnetic resonance image segmentation of the compressed spinal cord in patients with degenerative cervical myelopathy using convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 18(1), 45–54. <https://doi.org/10.1007/s11548-022-02783-0>
- Papinutto, N., Asteggiano, C., Bischof, A., Gundel, T. J., Caverzasi, E., Stern, W. A., Bastianello, S., Hauser, S. L., & Henry, R. G. (2020). Intersubject variability and normalization strategies for spinal cord total cross-sectional and gray matter areas. *Journal of Neuroimaging: Official Journal of the American Society of Neuroimaging*, 30(1), 110–118. <https://doi.org/10.1111/jon.12666>
- Prapas, I., Derakhshan, B., Mahdiraji, A. R., & Markl, V. (2021). Continuous training and deployment of deep learning models. *Datenbank-Spektrum: Zeitschrift Fur Datenbanktechnologie: Organ Der Fachgruppe Datenbanken Der Gesellschaft Fur Informatik e.V.*, 21(3), 203–212. <https://doi.org/10.1007/s13222-021-00386-8>
- Rahman, A., Ali, H., Badshah, N., Zakarya, M., Hussain, H., Rahman, I. U., Ahmed, A., & Haleem, M. (2022). Power mean based image segmentation in the presence of noise. *Scientific Reports*, 12(1), 21177. <https://doi.org/10.1038/s41598-022-25250-x>
- Rotem-Kohavi, N., Humphreys, S., Noonan, V. K., Cheng, C. L., Guay-Paquet, M., Bouthillier, M., Valošek, J., Naga Karthik, E., Lichtenstein, E., Guenther, N., Ost, K., Attabib, N., Hardisty, M., Badhiwala, J., Larouche, J., Pahuta, M., Christie, S., Fehlings, M. G., Fourney, D. ... Cadotte, D. W. (2025). Building a library of acute traumatic spinal cord injury images across Canada: A retrospective cohort study protocol. *BMJ Open*, 15(12), e106818. <https://doi.org/10.1136/bmjopen-2025-106818>
- Shi, J., & Wu, J. (2021). Distilling effective supervision for robust medical image segmentation with noisy labels. *arXiv [cs.CV]*. <http://arxiv.org/abs/2106.11099>
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755–759. <https://doi.org/10.1038/s41586-024-07566-y>
- SlicerNNUnet: 3D Slicer nnUNet integration to streamline usage for nnUNet based AI extensions. (n.d.). Github. Retrieved February 19, 2025, from <https://github.com/KitwareMedical/SlicerNNUnet>
- Sodhani, S., Faramarzi, M., Mehta, S. V., Malviya, P., Abdelsalam, M., Janarthanan, J., & Chandar, S. (2022). An Introduction to lifelong supervised learning. *arXiv [cs.LG]*. <http://arxiv.org/abs/2207.04354>
- Spijth, O., Frid, J., & Hellander, A. (2021). The machine learning life cycle and the cloud: Implications for drug discovery. *Expert Opinion on Drug Discovery*, 16(9), 1071–1079. <https://doi.org/10.1080/17460441.2021.1932812>
- Tabassam, A. I. U. (2023). MLOps: A step forward to enterprise machine learning. *arXiv [cs.SE]*. <http://arxiv.org/abs/2305.19298>
- Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, 15(1), 29. <https://doi.org/10.1186/s12880-015-0068-x>
- Taso, M., Girard, O. M., Duhamel, G., Le Troter, A., Feiweier, T., Guye, M., Ranjeva, J. P., & Callot, V. (2016). Tract-specific and age-related variations of the spinal cord microstructure: A multi-parametric MRI study using diffusion tensor imaging (DTI) and inhomogeneous magnetization transfer (ihMT). *NMR in Biomedicine*, 29(6), 817–832. <https://doi.org/10.1002/nbm.3530>
- Terpilowski, M. (2019). scikit-posthocs: Pairwise multiple comparison tests in Python. *Journal of Open Source Software*, 4(36), 1169. <https://doi.org/10.21105/joss.01169>
- Treveil, M., Omont, N., Stenac, C., Lefevre, K., Phan, D., Zentici, J., Lavoillotte, A., Miyazaki, M., & Heidmann, L. (2020). *Introducing MLOps*. O'Reilly Media. https://doi.org/10.1007/978-1-4842-9642-4_1
- Tsagkas, C., Horvath-Huck, A., Haas, T., Amann, M., Todea, A., Altermatt, A., Müller, J., Cagol, A., Leimbacher, M., Barakovic, M., Weigel, M., Pezold, S., Sprenger, T., Kappos, L., Bieri, O., Granziera, C., Cattin, P., & Parmar, K. (2023). Fully automatic method for reliable spinal cord compartment segmentation in multiple sclerosis. *AJNR. American Journal of Neuroradiology*, 44(2), 218–227. <https://doi.org/10.3174/ajnr.A7756>
- Valošek, J., Bédard, S., Keřkovský, M., Rohan, T., & Cohen-Adad, J. (2024). A database of the healthy human spinal cord morphometry in the PAM50 template space. *Imaging Neuroscience*, 2(34), 1–15. https://doi.org/10.1162/imag_a_00075
- Valošek, J., & Cohen-Adad, J. (2024). Reproducible spinal cord quantitative MRI analysis with the Spinal Cord Toolbox. *Magnetic Resonance in Medical Sciences: MRMS: An Official Journal of Japan Society of Magnetic Resonance in Medicine*, 23(3), 307–315. <https://doi.org/10.2463/mrms.rev.2023-0159>
- van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillion-Robin, J.-C., Pieper, S., & Aerts, H. J. W. L. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21), e104–e107. <https://doi.org/10.1158/0008-5472.CAN-17-0339>
- Veiga-Canuto, D., Cerdà-Alberich, L., Sangüesa Nebot, C., Martínez de Las Heras, B., Pötschger, U., Gabbelloni, M., Carot Sierra, J. M., Taschner-Mandl, S., Düster, V.,

- Cañete, A., Ladenstein, R., Neri, E., & Martí-Bonmatí, L. (2022). Comparative multicentric evaluation of inter-observer variability in manual and automatic segmentation of neuroblastic tumors in magnetic resonance images. *Cancers*, *14*(15), 3648. <https://doi.org/10.3390/cancers14153648>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Yao, J., Zhang, Y., Zheng, S., Goswami, M., Prasanna, P., & Chen, C. (2023). Learning to segment from noisy annotations: A spatial correction approach. *arXiv [eess.IV]*. <http://arxiv.org/abs/2308.02498>