Monitoring morphometric drift in lifelong learning segmentation of the spinal cord[1]

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, 1631 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 to 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 model is freely available in Spinal Cord Toolbox v7.0.

Keywords: Segmentation, MRI, Spinal Cord, MLOps, Lifelong Learning, Morphometric Drift

[1] Currently under review at the Imaging Neuroscience journal (https://direct.mit.edu/imag)

1. INTRODUCTION

Spinal cord segmentation is relevant for quantifying morphometric changes, such as cord atrophy in multiple sclerosis (MS) (Losseff et al. 1996; Lukas et al. 2013; Bautin and Cohen-Adad 2021), 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 and Cohen-Adad 2024).

While automatic tools for spinal cord segmentation exist, they have typically been developed in isolated, static environments (Masse-Gignac et al. 2023; Tsagkas et al. 2023; M. Chen et al. 2013; Nozawa et al. 2023; Gros et al. 2019; De Leener, Kadoury, and Cohen-Adad 2014). As a result, these tools rely on different procedures for creating ground truth (GT) masks, model architectures, and training strategies. Chen et al. (M. Chen et al. 2013) presented an atlas-based topology-preserving method for segmenting scans with different fields of view. Gros et al. (2019) proposed a collection of contrast-specific models (sct_deepseg_sc) trained on a multisite dataset of healthy controls and MS patients. However, the most commonly used variant is a convolutional network with 2D kernels, which prevents the models from capturing the full 3D spatial context and results in poor spinal cord segmentations in DCM and SCI patients with lesions. Masse-Gignac et al. (2023) proposed a cascade of two Convolutional Neural Networks (CNNs), trained separately on axial and sagittal T2w scans, for segmenting injured spinal cords. The GT masks used for training were adapted from the segmentations obtained initially by sct_deepseg_sc 2D. Nozawa et al. (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 public dataset of healthy participants (Cohen-Adad et al. 2021b), which generalizes across contrasts but struggles to segment the cord in pathological cases such as DCM as a consequence of being trained only on healthy participants' data. SCIseg (Naga Karthik, Valošek, et al. 2025), a 3D model trained exclusively on T2-weighted images for the segmentation of the spinal cord and intramedullary lesions in DCM and SCI patients, improves segmentation in pathological cases but is limited by its reliance on a single contrast. With the plethora of models specializing in a specific set of contrasts and pathologies, there is a lack of standardization across all phases of developing an automatic segmentation pipeline. Furthermore, no continuous learning pipeline is in place to monitor the drift/degradation in the segmentation performance of these models over time.

Morphometric measures derived from spinal cord segmentations are highly dependent on the method used (Cohen-Adad et al. 2021b; Bédard et al. 2025) and may drift as the methods are iterated upon. This can result in a discrepancy between normative values of healthy populations evaluated by each segmentation method. In addition, morphometric measures suffer from large inter-subject variability due to factors such as sex and age, limiting our ability to detect subtle morphometric changes (Valošek et al. 2024; Labounek et al. 2024; Papinutto et al. 2020; Bédard and Cohen-Adad 2022; Bédard et al. 2024; Taso et al. 2016). One approach to mitigate this variability is to compare them with morphometrics obtained from healthy controls (Valošek et al. 2024; Labounek et al. 2024; Bédard et al. 2024; Horáková et al. 2022; Kato et al. 2012). These normalization techniques assume that the morphometrics of new subjects are computed using the same method as the original normative database (Valošek et al. 2024). However, this is an assumption which 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[2] framework (Prapas et al. 2021; Agirre, Jonsson, and Larcher 2021; Liu and Mazumder 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 9 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.

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.

[2] Note that this differs from the classical approach to lifelong/continual learning, where it is assumed that access to previously available data is constrained. Instead, we focus on the *post-deployment* aspect of the learning system which is continuously updated over time with unrestrained access to data.

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 and training protocol

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 volumes[3]). 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 mm to 1 mm), anisotropic axially-oriented (in-plane resolution: 0.29 mm to 1 mm; slice thickness: 1 mm to 9.3 mm) and sagittally-oriented (in-plane resolution: 0.28 mm to 1 mm; slice thickness: 0.8 mm to 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 **Table S1** provides more details on the distribution of image resolutions for each contrast.

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

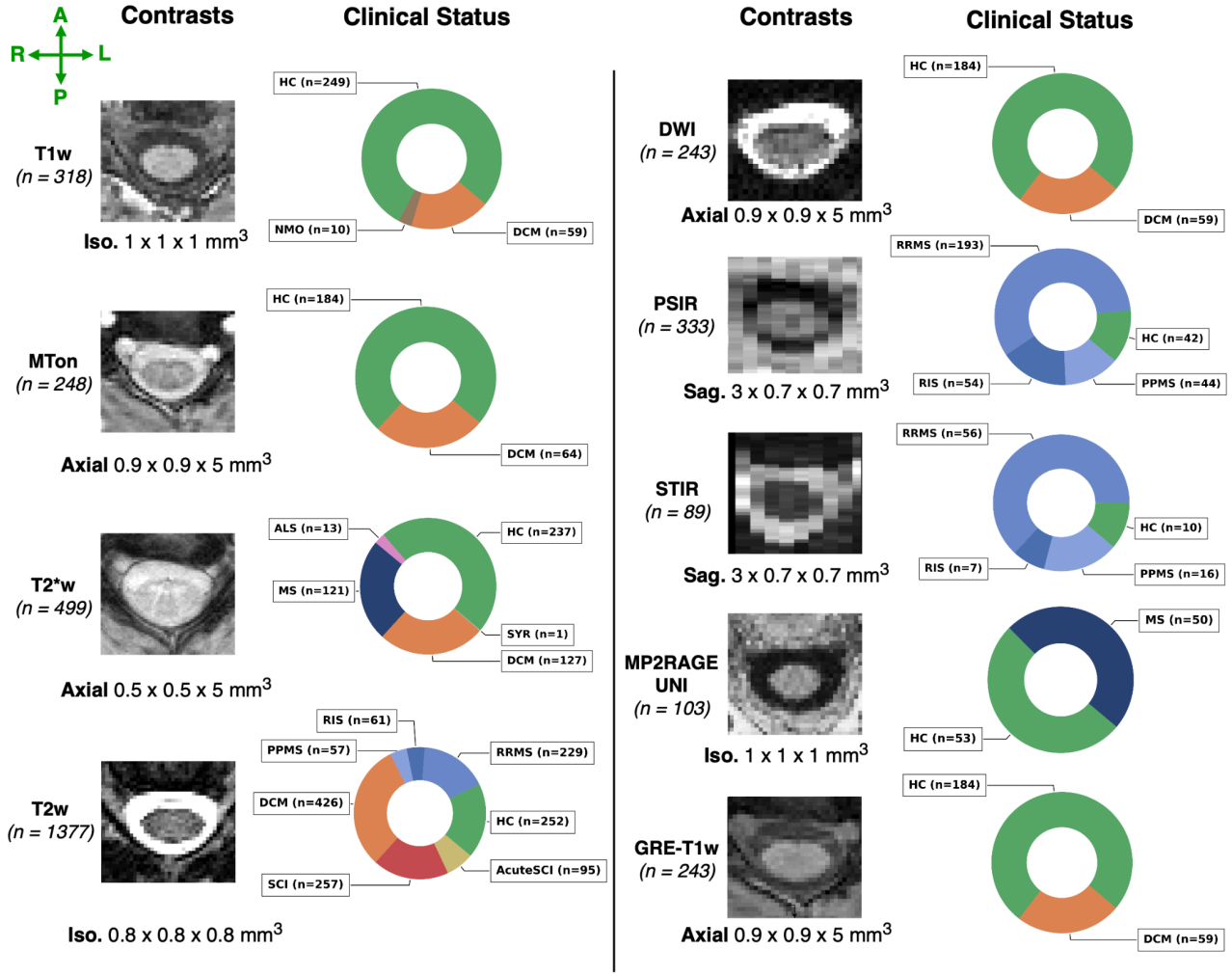

Figure 1. Overview of the dataset and image characteristics*. Representative axial slices of 9 contrasts and the total of images used for each contrast in brackets, the orientation (axial/sagittal) along with the median resolution of images. The respective doughnut chart illustrates the proportion of 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.*

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 as described in (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 and Cohen-Adad 2024). Four experienced raters (ENK, SB, JV, JCA) 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 (Naga Karthik, Valošek, et al. 2025; Karthik et al. 2024) was used instead, followed by manual corrections by JV and ENK. 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. 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 subject-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 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 subject and each site, the aggregated dataset included 2,945 training and 508 testing images.

2.1.3. Data preprocessing, augmentation and training

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 x 0.7 x 1] mm^3 and the median shape was [96 x 320 x 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 x 224 x 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. Lastly, all images were normalized using *z*-score normalization.

The network architecture is a standard convolutional neural network architecture 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→64→128→256→320→320). The network was trained with a combination of Dice (Milletari, Navab, and Ahmadi 2016) and cross-entropy losses. At each layer, 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 model was trained using 5-fold cross-validation for 1000 epochs, a batch size of two, and with the stochastic gradient descent optimizer and a polynomial learning rate scheduler. All experiments were run on a single 48 GB NVIDIA A6000 GPU.

2.2. Lifelong learning for morphometric drift monitoring

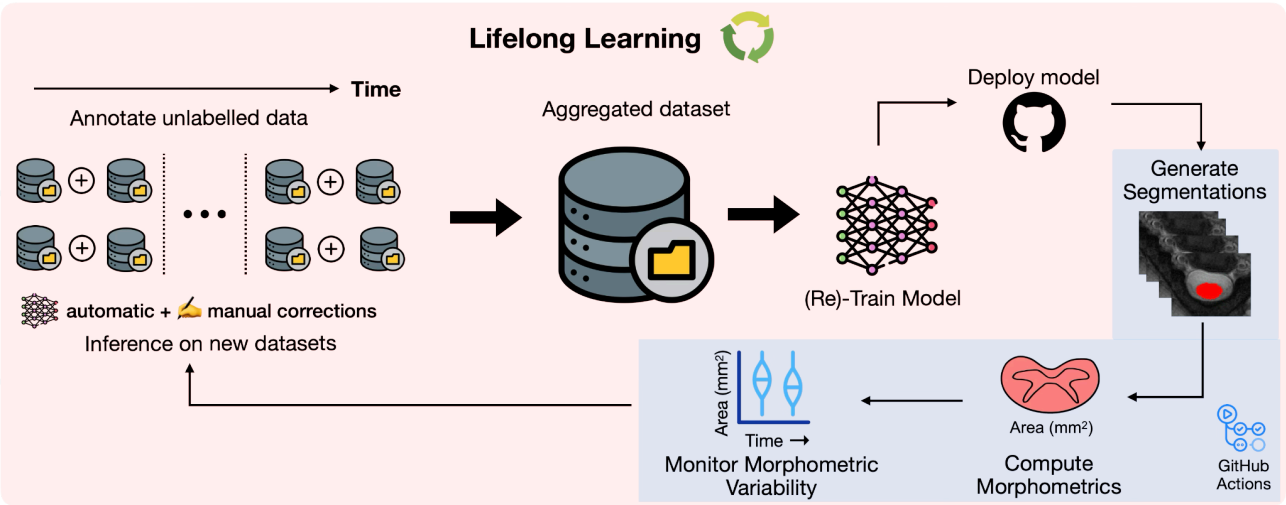

Figure 2. Overview of the lifelong learning strategy for continuous training of spinal cord segmentation models. *Unlabelled data containing various contrasts and pathologies, gathered from multiple sites worldwide, are segmented automatically with an existing state-of-the-art model and undergo visual quality control for inconsistencies in segmentations, excluding data with artifacts. Labelled 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.*

We take an *MLOps* (Tabassam 2023; Alla and Adari 2020; Treveil et al. 2020) approach to propose our lifelong learning framework for monitoring morphometric drift across various versions of the model (**Figure 2**). Once the segmentation model is trained, we deploy the model as an official

release on GitHub. 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 **Figure 3** for pseudocode of the workflow). To ensure a fair comparison with previous work (Bédard et al. 2025; Gros et al. 2019), the spinal cord cross-sectional area (CSA) is computed on a frozen test of healthy participants (*n=49*) containing 6 contrasts (T1w, T2w, T2*w, DWI, MT-on, GRE-T1w) for each participant. 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 unlabelled datasets (arriving in the future), perform QC, add them to growing collection of datasets, and 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).

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
```

Figure 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.*

2.3. Validation protocol

2.3.1. Evaluation Metrics

To evaluate the segmentation accuracy quantitatively, we report the Dice coefficient, average surface distance (ASD), and relative volume error (RVE) on the 'frozen' test mentioned previously. For a more clinically oriented assessment of the models, we also computed CSA averaged over C2-C3 vertebral levels of the cervical spinal cord on the same test set predictions to measure the morphometric variability for each model. These measurements are done 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.

2.3.2. Qualitative evaluation of segmentations on various contrasts and pathologies

We compared the segmentations between our model and previous state-of-the-art models 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.

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 Dice score, Relative Volume Error (RVE) and Surface Distance, from the ANIMA toolbox (Commowick et al. 2018).

2.3.3. 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).

2.3.4. 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.

2.3.5. 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 the *Introduction*, 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 6 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.

$$Scaling\ Factor\ = \frac{metric_{contrast-agnostic_v3.0}}{metric_{sct_deepseg_sc}}$$

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).

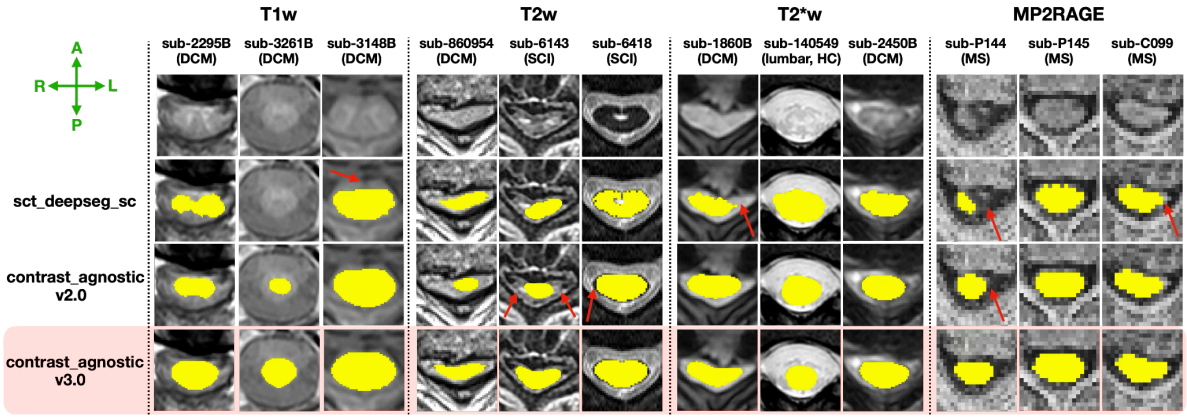

Figure 4. *Comparison of the automatic spinal cord segmentations between `contrast_agnostic_v3.0` (current version, highlighted), `contrast_agnostic_v2.0` and `sct_deepseg_sc` on healthy controls (HC), DCM, SCI and MS patients on the test set (unseen during training). Red arrows show the instances where the previous models fail, particularly under heavy compression (with/without lesions) in sub-860594, sub-6143 and sub-1860B.*

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 subject IDs in the bottom left). Notably, in the case of whole-spine images, the model learned to segment the entire spine despite only being trained independently on individual cervical, thoracic and lumbar segments.

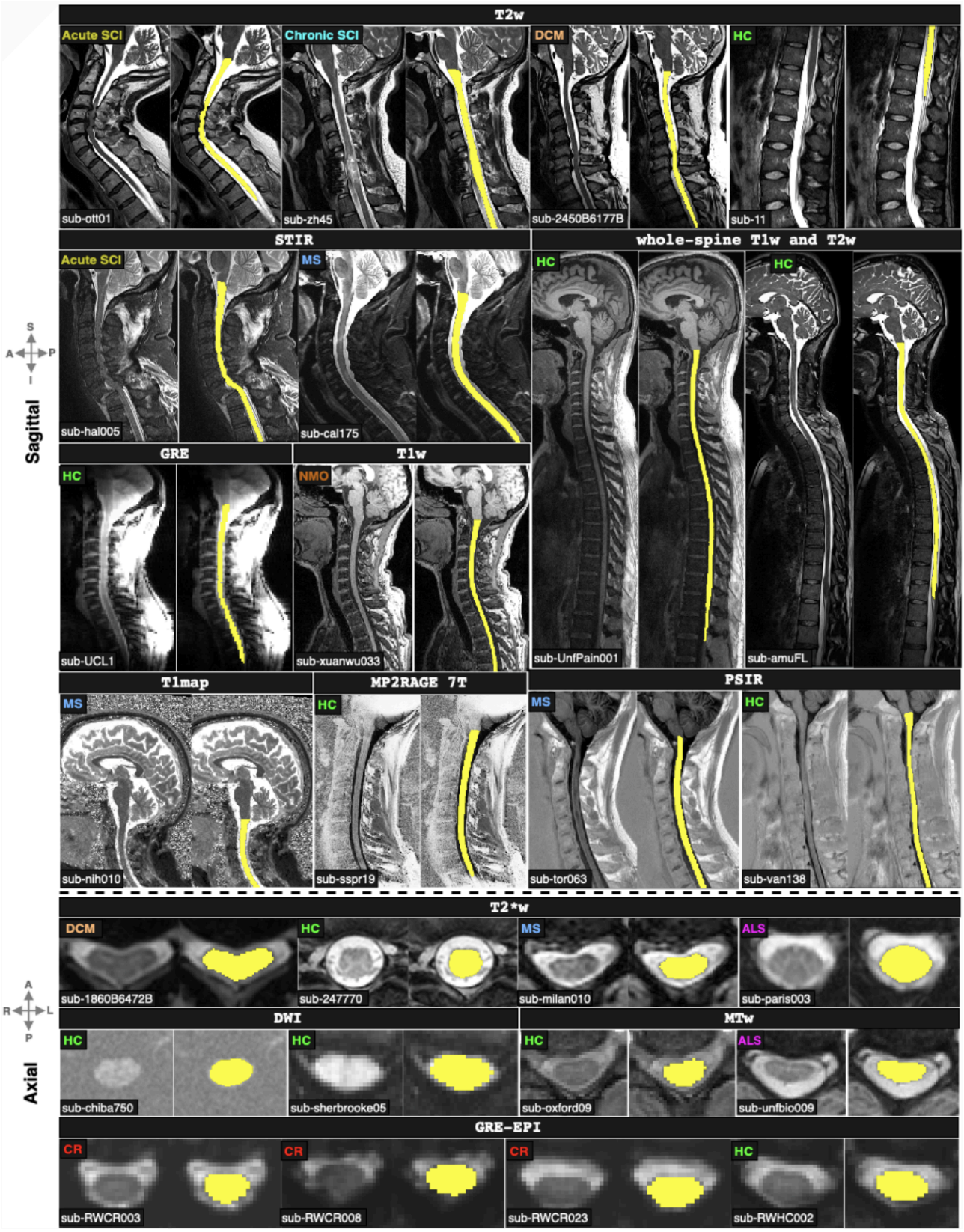
T2w
Acute SCI
sub-ott01
Chronic SCI
sub-zh45
DCM
sub-2450B6177B
HC
sub-11
STIR
Acute SCI
sub-hal005
MS
sub-cal175
whole-spine T1w and T2w
HC
sub-UnfPain001
HC
sub-amuFL
GRE
HC
sub-UCL1
T1w
NMO
sub-xuanwu033
T1map
MS
sub-nih010
MP2RAGE 7T
HC
sub-sspr19
PSIR
MS
sub-tor063
HC
sub-van138
Sagittal
S
A
P
I
T2*w
DCM
sub-1860B6472B
HC
sub-247770
MS
sub-milan010
ALS
sub-paris003
DWI
HC
sub-chiba750
HC
sub-sherbrooke05
MTw
HC
sub-oxford09
ALS
sub-unfbio009
GRE-EPI
CR
sub-RWCR003
CR
sub-RWCR008
CR
sub-RWCR023
HC
sub-RWHC002
Axial
A
R
L
P

Figure 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.*

3.1.2. Quantitative evaluation on healthy controls and pathologies

Table 1 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 1A)**, we then present results for test sets containing T2*w images of MS patients from two sites (**Table 1B**), axial and sagittal T2w scans of DCM patients from two sites (**Table 1D**), and axial and sagittal T2w scans of traumatic SCI (acute, intermediate and chronic phases) from six sites (**Table 1D**). In all comparisons, the proposed `contrast_agnostic_v3.0` model achieved similar or better performance compared to the previous state-of-the-art or pathology-specific models.

Table 1. *Quantitative comparison of spinal cord segmentations for previous segmentation methods on the test set (n = 49 participants; n_{vol} = 294 images) averaged across all contrasts. Quantitative comparison on patients with MS on T2*w contrast (n = 36 participants; n_{vol} = 36 images). Quantitative comparison on patients with DCM on axial and sagittal T2w scans (n_{vol} = 39) RVE stands for Relative Volume Error, and ASD stands for Average Surface Distance.*

Methods	**Dice (↑)**	**RVE (%)**	**ASD (↓)**
	Opt. value: 1	Opt. value: 0	Opt. value: 0
A) Healthy participants (n = 49; 6 contrasts per participant; n_{vol} = 294)			
sct_deepseg_sc	0.95 ± 0.03	-0.18 ± 8.95	0.04 ± 0.27
contrast_agnostic_v2.0	0.95 ± 0.02	**-0.05 ± 4.18**	**0.02 ± 0.12**
contrast_agnostic_v3.0 (proposed)	**0.96 ± 0.02**	-0.76 ± 4.59	0.04 ± 0.27

B) Patients with MS (n = 36; T2*w contrast; n_{vol} = 36)			
sct_deepseg_sc	0.94 ± 0.02	-9.03 ± 3.35	**0.003 ± 0.009**
contrast_agnostic_v2.0	0.94 ± 0.01	-10.12 ± 2.89	0.009 ± 0.016
contrast_agnostic_v3.0 (proposed)	**0.96 ± 0.01**	**-5.34 ± 2.89**	0.005 ± 0.014
C) Patients with DCM (n = 39; T2w contrast; n_{vol} = 39)			
SCIsegV2	**0.97 ± 0.01**	**-2.34 ± 1.79**	0.001 ± 0.001
contrast_agnostic_v2.0	0.91 ± 0.02	-11.91 ± 4.16	0.01 ± 0.04
contrast_agnostic_v3.0 (proposed)	0.96 ± 0.01	-2.51 ± 2.25	**0.001 ± 0.001**
D) Patients with SCI (n = 60; T2w contrast; n_{vol} = 60)			
SCIsegV2	**0.93 ± 0.04**	5.22 ± 7.63	**0.01 ± 0.01**
sct_deepseg_sc	0.82 ± 0.23	-13.68 ± 24.1	7.61 ± 31.87
contrast_agnostic_v2.0	0.74 ± 0.17	-28.81 ± 20.49	1.38 ± 4.56
contrast_agnostic_v3.0 (proposed)	**0.93 ± 0.06**	**1.75 ± 14.63**	0.01 ± 0.04

3.2. Quantitative evaluation of morphometric drift across model versions

3.2.1. Variability of CSA across contrasts

The figures below are automatically output 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) `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 to the other methods. **Figure S1** plots the variability in spinal cord CSA per each

individual contrast. Similar to the analysis of CSA variability 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 **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 `contrast-agnostic_v2.0` was trained only on a healthy participant database. **Table S2** compares the models' performances using the Dice score, RVE and ASD metrics.

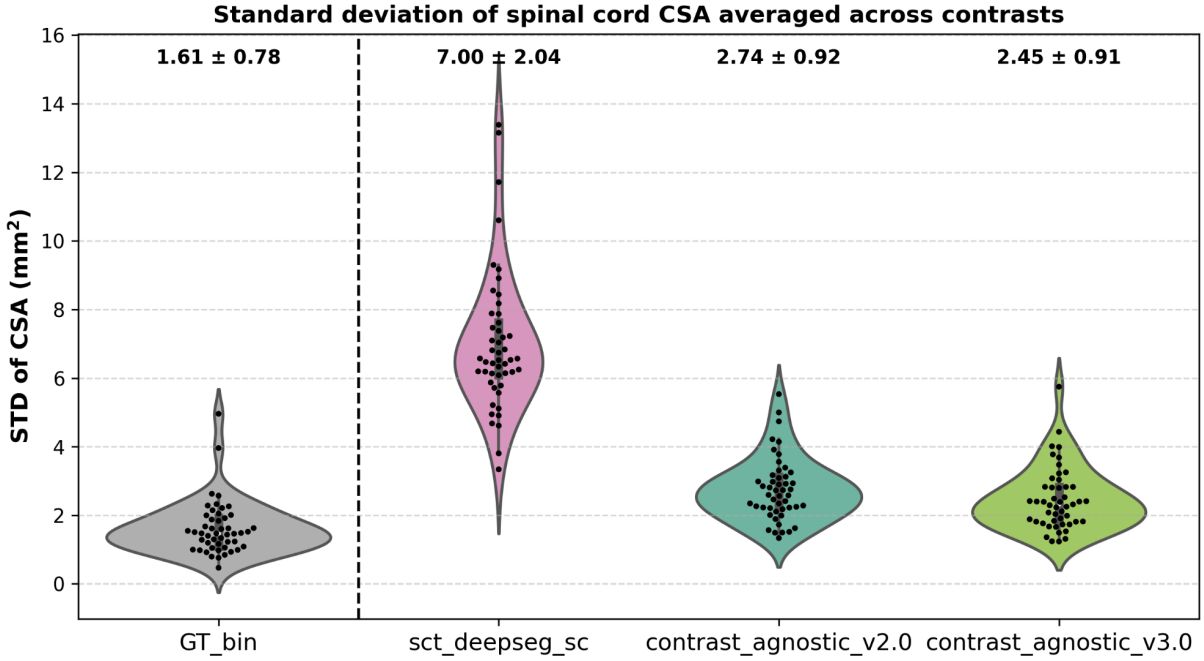

Figure 6. *CSA variability measured in terms of the standard deviation across 6 contrasts on a test set of healthy participants (n=49). Our proposed model achieved the lowest STD averaged across 6 contrasts (i.e. each point shows the mean of 6 contrasts for the given participant) showing more stability in segmentations across contrasts. The lower the CSA STD across contrasts, the better.*

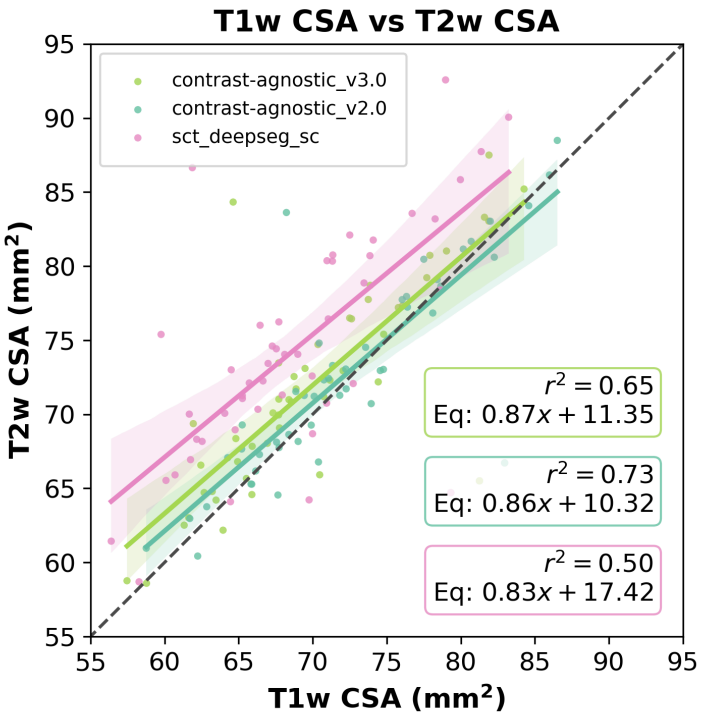

Figure 7. *Level of agreement between T1w and T2w CSA at C2-C3 for `contrast_agnostic_v3.0`, `contrast_agnostic_v2.0` and `sct_deepseg_sc`. Each point represents one participant. The black dashed line represents perfect agreement between the CSA of T1w and T2w contrasts.*

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 performed 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 to 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 recurisely generated GT masks underestimated the CSA on all contrasts.

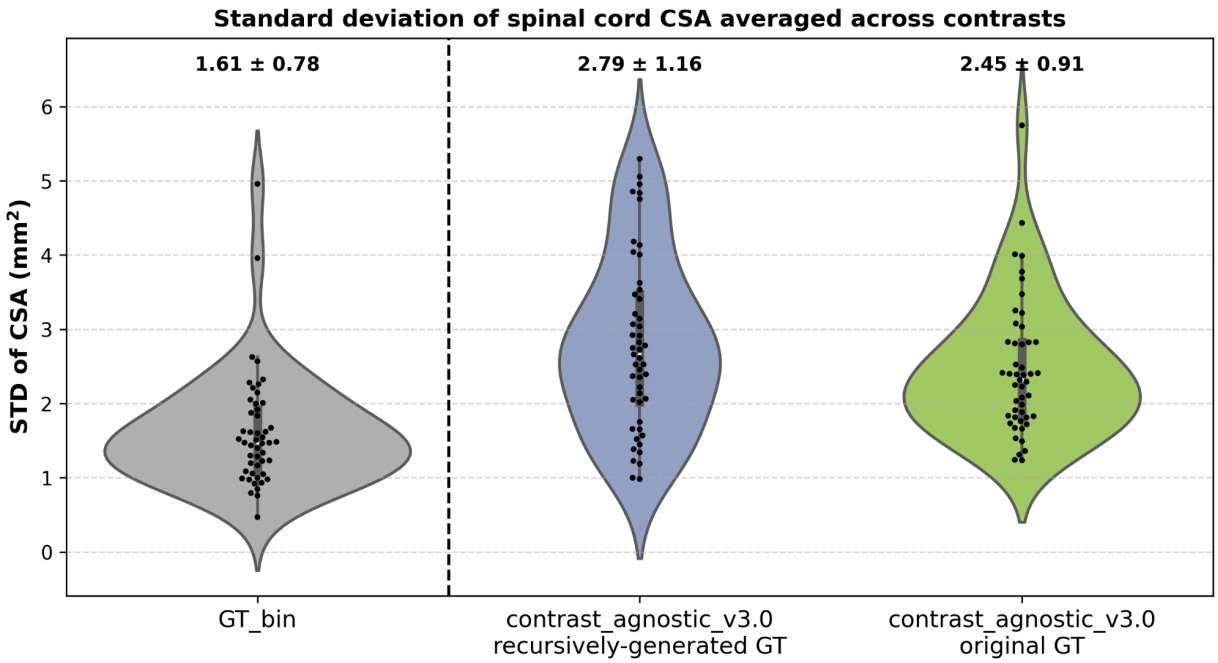

Figure 8. *Standard deviation of the CSA across 6 contrasts for models trained on: (i) recursively generated GT masks (blue), and (ii) original GT masks (green). Each point shows the mean of 6 contrasts for the given participant. The model trained on noisy labels tends to produce stable segmentations resulting in a lower STD across contrasts. The lower the CSA STD across contrasts, the better.*

3.2.3. Normative database results

Figure 9A shows the plots for 6 different morphometric measures computed on 203 healthy participants using two 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 computed with the 2 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] https://github.com/spinalcordtoolbox/PAM50-normalized-metrics/releases/tag/r20250321

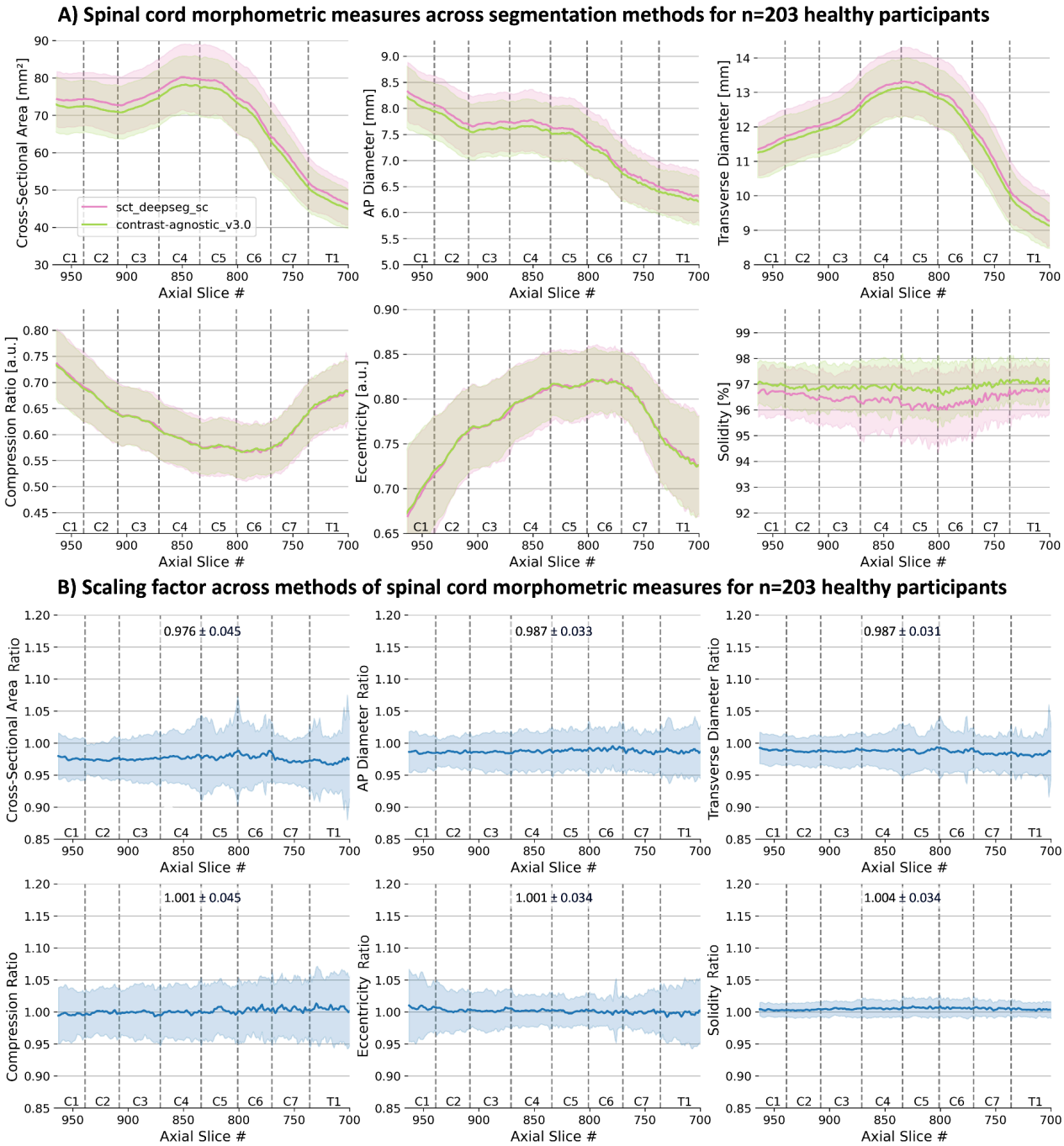

Figure 9. ***(A)*** *Morphometric measures computed on n=203 healthy participants from the Spine Generic Dataset* ***(Cohen-Adad et al. 2021b)*** *for 6 morphometric measures using 2 different segmentation methods:* `sct_deepseg_sc` *with manual correction (green) and* `contrast-agnostic_v3.0` *(orange) with* ***(B)*** *scaling factor between the methods means ± std are displayed. Metrics are shown in the PAM50 space.*

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 tends 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 (Yao et al. 2023; Shi and Wu 2021). However, noise in training data tends to disrupt model training by making the models unintentionally focus on such outliers (Rahman et al. 2022; Taha and 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 labelled each new dataset containing new contrasts or pathologies with existing automatic models (Gros et al. 2019; Bédard et al. 2025; 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 a compressed snapshot of the dataset, which is useful for sharing with the clinical sites (Jwa, Norgaard, and Poldrack 2025).

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 **Figure 4** and **Figure 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). Interestingly, 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 to existing pathology-specific models (**Table 1**) 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, Jonsson, and Larcher 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) (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 (Valošek et al. 2024; Joo et al. 2025). 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) unlabelled data to train subsequent models. 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, Jonsson, and Larcher 2021; Prapas et al. 2021; González et al. 2024), 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 arrives, 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 (**Figure 8**, **Figure S3**), we observed that the model underestimated the average CSA on a healthy subset of participants for each of the 6 contrasts and resulted in a higher CSA STD across contrasts, when compared to 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. On the other hand, 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.

4.2.4. Binary vs. 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, subtly 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. At these regions, soft masks can better represent the partial volume compared to binary masks.

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 6 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 on participants with pathologies as well.

Another issue is the stagnation of the training data distribution when developing models over time. With subsequent models being trained on new data (potentially from different populations –

[5] https://portal.fli-iam.irisa.fr/ms-multi-spine/

[6] https://ivdm3seg.weebly.com

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.

7. REFERENCES

Agirre, Eneko, Anders Jonsson, and Anthony Larcher. 2021. "Framing Lifelong Learning as Autonomous Deployment: Tune Once Live Forever." In *Lecture Notes in Electrical Engineering*, 0:331–36. Lecture Notes in Electrical Engineering. Singapore: Springer Singapore.

Alla, Sridhar, and Suman Kalyan Adari. 2020. *Beginning MLOps with MLFlow: Deploy Models in AWS SageMaker, Google Cloud, and Microsoft Azure*. 1st ed. Berlin, Germany: APress.

Bautin, Paul, and Julien Cohen-Adad. 2021. "Minimum Detectable Spinal Cord Atrophy with Automatic Segmentation: Investigations Using an Open-Access Dataset of Healthy Participants." *NeuroImage. Clinical* 32 (102849): 102849.

Bédard, Sandrine, and Julien Cohen-Adad. 2022. "Automatic Measure and Normalization of Spinal Cord Cross-Sectional Area Using the Pontomedullary Junction." *Frontiers in Neuroimaging* 1 (November):1031253.

Bédard, Sandrine, Enamundram Naga Karthik, Charidimos Tsagkas, Emanuele Pravatà, Cristina Granziera, Andrew Smith, Kenneth Arnold Weber Ii, and Julien Cohen-Adad. 2025. "Towards Contrast-Agnostic Soft Segmentation of the Spinal Cord." *Medical Image Analysis* 101 (103473): 103473.

Bédard, Sandrine, Jan Valošek, Maryam Seif, Armin Curt, Simon Schading, Nikolai Pfender, Patrick Freund, Markus Hupp, and Julien Cohen-Adad. 2024. "Normalizing Spinal Cord Compression Morphometric Measures: Application in Degenerative Cervical Myelopathy." *medRxiv*. https://doi.org/10.1101/2024.03.13.24304177.

Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. "Rethinking Atrous Convolution for Semantic Image Segmentation." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1706.05587.

Chen, Min, Aaron Carass, Jiwon Oh, Govind Nair, Dzung L. Pham, Daniel S. Reich, and Jerry L. Prince. 2013. "Automatic Magnetic Resonance Spinal Cord Segmentation with Topology Constraints for Variable Fields of View." *NeuroImage* 83 (December):1051–62.

Cohen-Adad, Julien, Eva Alonso-Ortiz, Mihael Abramovic, Carina Arneitz, Nicole Atcheson, Laura Barlow, Robert L. Barry, et al. 2021a. "Generic Acquisition Protocol for Quantitative MRI of the Spinal Cord." *Nature Protocols* 16 (10): 4611–32.

———. 2021b. "Open-Access Quantitative MRI Data of the Spinal Cord and Reproducibility across Participants, Sites and Manufacturers." *Scientific Data* 8 (1): 219.

Commowick, Olivier, Audrey Istace, Michaël Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, et al. 2018. "Objective Evaluation of Multiple Sclerosis Lesion Segmentation Using a Data Management and Processing Infrastructure." *Scientific Reports* 8 (1): 13650.

De Leener, Benjamin, Samuel Kadoury, and Julien Cohen-Adad. 2014. "Robust, Accurate and Fast Automatic Segmentation of the Spinal Cord." *NeuroImage* 98 (September):528–36.

De Leener, Benjamin, Simon Lévy, Sara M. Dupont, Vladimir S. Fonov, Nikola Stikov, D. Louis Collins, Virginie Callot, and Julien Cohen-Adad. 2017. "SCT: Spinal Cord Toolbox, an Open-Source Software for Processing Spinal Cord MRI Data." *NeuroImage* 145 (Pt A): 24–43.

Dou, Qi, Lequan Yu, Hao Chen, Yueming Jin, Xin Yang, Jing Qin, and Pheng-Ann Heng. 2017. "3D Deeply Supervised Network for Automated Segmentation of Volumetric Medical Images." *Medical Image Analysis* 41 (October):40–54.

González, Camila, Moritz Fuchs, Daniel Pinto dos Santos, Philipp Matthies, Manuel Trenz, Maximilian Grüning, Akshay Chaudhari, et al. 2024. "Regulating Radiology AI Medical Devices

That Evolve in Their Lifecycle." *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2412.20498.
Griethuysen, Joost J. M. van, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G. H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J. W. L. Aerts. 2017. "Computational Radiomics System to Decode the Radiographic Phenotype." *Cancer Research* 77 (21): e104–7.
Gros, Charley, Benjamin De Leener, Atef Badji, Josefina Maranzano, Dominique Eden, Sara M. Dupont, Jason Talbott, et al. 2019. "Automatic Segmentation of the Spinal Cord and Intramedullary Multiple Sclerosis Lesions with Convolutional Neural Networks." *NeuroImage* 184 (January):901–15.
Horáková, Magda, Tomáš Horák, Jan Valošek, Tomáš Rohan, Eva Koriťáková, Marek Dostál, Jan Kočica, et al. 2022. "Semi-Automated Detection of Cervical Spinal Cord Compression with the Spinal Cord Toolbox." *Quantitative Imaging in Medicine and Surgery* 12 (4): 2261–79.
Isensee, Fabian, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. 2021. "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation." *Nature Methods* 18 (2): 203–11.
Isensee, Fabian, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jaeger. 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, Bio, Hyung Jun Park, Mina Park, Sang Hyun Suh, and Sung Jun Ahn. 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): 146–55.
Jwa, Anita S., Martin Norgaard, and Russell A. Poldrack. 2025. "Can I Have Your Data? Recommendations and Practical Tips for Sharing Neuroimaging Data upon a Direct Personal Request." *Imaging Neuroscience* 3 (March). https://doi.org/10.1162/imag_a_00508.
Karthik, Enamundram Naga, Anne Kerbrat, Pierre Labauge, Tobias Granberg, Jason Talbott, Daniel S. Reich, Massimo Filippi, et al. 2022. "Segmentation of Multiple Sclerosis Lesions across Hospitals: Learn Continually or Train from Scratch?" *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2210.15091.
Karthik, Enamundram Naga, Jan Valošek, Lynn Farner, Dario Pfyffer, Simon Schading-Sassenhausen, Anna Lebret, Gergely David, et al. 2024. "SCIsegV2: A Universal Tool for Segmentation of Intramedullary Lesions in Spinal Cord Injury." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2407.17265.
Kato, F., Y. Yukawa, K. Suda, M. Yamagata, and T. Ueta. 2012. "Normal Morphology, Age-Related Changes and Abnormal Findings of the Cervical Spine. Part II: Magnetic Resonance Imaging of over 1,200 Asymptomatic …." *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*. https://link.springer.com/article/10.1007/s00586-012-2176-4.
Labounek, René, Monica T. Bondy, Amy L. Paulson, Sandrine Bédard, Mihael Abramovic, Eva Alonso-Ortiz, Nicole T. Atcheson, et al. 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, Andreanne, Charley Gros, Enamundram Naga Karthik, and Julien Cohen-Adad. 2022. "Label Fusion and Training Methods for Reliable Representation of Inter-Rater Uncertainty." *arXiv [eess.IV]*. arXiv. http://arxiv.org/abs/2202.07550.
Liu, Bing, and Sahisnu Mazumder. 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): 15058–63.

Losseff, N. A., S. L. Webb, J. I. O'Riordan, R. Page, L. Wang, G. J. Barker, P. S. Tofts, W. I. McDonald, D. H. Miller, and A. J. Thompson. 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) (June):701–8.

Lukas, Carsten, Madeleine H. Sombekke, Barbara Bellenberg, Horst K. Hahn, Veronica Popescu, Kerstin Bendfeldt, Ernst W. Radue, et al. 2013. "Relevance of Spinal Cord Abnormalities to Clinical Disability in Multiple Sclerosis: MR Imaging Findings in a Large Cohort of Patients." *Radiology* 269 (2): 542–52.

Martin, Allan R., Benjamin De Leener, Julien Cohen-Adad, Sukhvinder Kalsi-Ryan, David W. Cadotte, Jefferson R. Wilson, Lindsay Tetreault, et al. 2018. "Monitoring for Myelopathic Progression with Multiparametric Quantitative MRI." Edited by Mathias Toft. *PloS One* 13 (4): e0195733.

Masse-Gignac, Nicolas, Salomón Flórez-Jiménez, Jean-Marc Mac-Thiong, and Luc Duong. 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.

Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1606.04797.

Molinier, Nathan, Sandrine Bédard, Mathieu Boudreau, Julien Cohen-Adad, Virginie Callot, Eva Alonso-Ortiz, Charles Pageot, and Nilser Laines-Medina. 2024. "'Whole-Spine.'" OpenNeuro. https://doi.org/10.18112/openneuro.ds005616.v1.0.1.

Naga Karthik, Enamundram, Julian McGinnis, Ricarda Wurm, Sebastian Ruehling, Robert Graf, Jan Valosek, Pierre-Louis Benveniste, et al. 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, Enamundram, Jan Valošek, Andrew C. Smith, Dario Pfyffer, Simon Schading-Sassenhausen, Lynn Farner, Kenneth A. Weber 2nd, Patrick Freund, and Julien Cohen-Adad. 2025. "SCIseg: Automatic Segmentation of Intramedullary Lesions in Spinal Cord Injury on T2-Weighted MRI Scans." *Radiology. Artificial Intelligence* 7 (1): e240005.

Nozawa, Kyohei, Satoshi Maki, Takeo Furuya, Sho Okimatsu, Takaki Inoue, Atsushi Yunde, Masataka Miura, et al. 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.

Papinutto, Nico, Carlo Asteggiano, Antje Bischof, Tristan J. Gundel, Eduardo Caverzasi, William A. Stern, Stefano Bastianello, Stephen L. Hauser, and Roland G. Henry. 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): 110–18.

Prapas, Ioannis, Behrouz Derakhshan, Alireza Rezaei Mahdiraji, and Volker Markl. 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): 203–12.

Rahman, Afzal, Haider Ali, Noor Badshah, Muhammad Zakarya, Hameed Hussain, Izaz Ur Rahman, Aftab Ahmed, and Muhammad Haleem. 2022. "Power Mean Based Image

Segmentation in the Presence of Noise." *Scientific Reports* 12 (1): 21177.
Shi, Jialin, and Ji Wu. 2021. "Distilling Effective Supervision for Robust Medical Image Segmentation with Noisy Labels." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2106.11099.
Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. "AI Models Collapse When Trained on Recursively Generated Data." *Nature* 631 (8022): 755–59.
SlicerNNUnet: 3D Slicer nnUNet Integration to Streamline Usage for nnUNet Based AI Extensions. n.d. Github. Accessed February 19, 2025. https://github.com/KitwareMedical/SlicerNNUnet.
Sodhani, Shagun, Mojtaba Faramarzi, Sanket Vaibhav Mehta, Pranshu Malviya, Mohamed Abdelsalam, Janarthanan Janarthanan, and Sarath Chandar. 2022. "An Introduction to Lifelong Supervised Learning." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2207.04354.
Tabassam, A. I. Ullah. 2023. "MLOps: A Step Forward to Enterprise Machine Learning." *arXiv [cs.SE]*. arXiv. http://arxiv.org/abs/2305.19298.
Taha, Abdel Aziz, and Allan Hanbury. 2015. "Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool." *BMC Medical Imaging* 15 (1): 29.
Taso, Manuel, Olivier M. Girard, Guillaume Duhamel, Arnaud Le Troter, Thorsten Feiweier, Maxime Guye, Jean Philippe Ranjeva, and Virginie Callot. 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): 817–32.
Treveil, Mark, Nicolas Omont, Clement Stenac, Kenji Lefevre, Du Phan, Joachim Zentici, Adrien Lavoillotte, Makoto Miyazaki, and Lynn Heidmann. 2020. *Introducing MLOps*. Sebastopol, CA: O'Reilly Media.
Tsagkas, C., A. Horvath-Huck, T. Haas, M. Amann, A. Todea, A. Altermatt, J. Müller, et al. 2023. "Fully Automatic Method for Reliable Spinal Cord Compartment Segmentation in Multiple Sclerosis." *AJNR. American Journal of Neuroradiology* 44 (2): 218–27.
Valošek, Jan, Sandrine Bédard, Miloš Keřkovský, Tomáš Rohan, and Julien Cohen-Adad. 2024. "A Database of the Healthy Human Spinal Cord Morphometry in the PAM50 Template Space." *Imaging Neuroscience* 2 (34): 1–15.
Valošek, Jan, and Julien Cohen-Adad. 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): 307–15.
Yao, Jiachen, Yikai Zhang, Songzhu Zheng, Mayank Goswami, Prateek Prasanna, and Chao Chen. 2023. "Learning to Segment from Noisy Annotations: A Spatial Correction Approach." *arXiv [eess.IV]*. arXiv. http://arxiv.org/abs/2308.02498.

Supplementary Material: Minimizing morphometric drift in lifelong learning segmentation of the spinal cord

This document contains the supplementary material presenting an overview of the dataset characteristics, additional plots comparing the variability in spinal cord cross-sectional area (CSA) per individual contrasts, CSA variability split across vendors, and quantitative comparison in terms of the Dice scores, relative volume error and average surface distance.

S1.1. Dataset characteristics

Table S1 contains a detailed overview of the range of image resolutions and orientations for each contrast in the dataset. Note that the images span a wide range of resolutions, especially with thick slices in the axial orientation for a few contrasts.

Table S1. *Dataset characteristics grouped by image orientation (axial, sagittal) and resolution (isotropic, anisotropic) for each contrast. Mean in-plane resolution and mean slice thickness are shown, followed by their respective minimum and maximum range of resolutions (in [ ]).*

Contrasts	**Isotropic**		**Anisotropic Axial Orientation**		**Anisotropic Sagittal Orientation**	
	in-plane resolution (mm^2)	slice thickness (mm)	in-plane resolution (mm^2)	slice thickness (mm)	in-plane resolution (mm^2)	slice thickness (mm)
T1-w	1.0 x 1.0 [1.0, 1.0]	1.0 [1.0, 1.0]	0.35 x 0.35 [0.35 x 0.35, 0.35 x 0.35]	2.54 [2.5, 5.0]	1.0 x 1.0 [1.0, 1.0]	1.0 [1.0, 1.0]
T2-w	0.8 x 0.8 [0.8, 0.8]	0.8 [0.8, 0.8]	0.5 x 0.5 [0.3 x 0.3, 0.8 x 1.0]	3.8 [1.0, 7.0]	0.48 x 0.48 [0.28 x 0.28, 0.96 x 0.96]	2.13 [0.8, 4.83]
T2*-w	–	–	0.44 x 0.44 [0.29 x 0.29, 0.5 x 0.5]	4.93 [2.5, 9.2]	–	–

MT-on	–	–	0.89 x 0.89 [0.62 x 0.62, 0.9 x 0.9]	5.06 [5.0, 9.3]	–	–
GRE-T1w	–	–	0.89 x 0.89 [0.68 x 0.68, 0.9 x 0.9]	5.0 [5.0, 5.0]	–	–
DWI	–	–	0.89 x 0.89 [0.34 x 0.34, 1.0 x 1.0]	5.0 [4.91, 5.0]	–	–
PSIR	–	–	–	–	0.69 x 0.69 [0.67 x 0.67, 0.69 x 0.69]	3.0 [3.0, 3.0]
STIR	–	–	–	–	0.7 x 0.7 [0.7 x 0.7, 0.7 x 0.7]	3.0 [3.0, 3.0]
MP2RAGE UNIT1	1.0 x 1.0 [1.0, 1.0]	1.0 [1.0, 1.0]	–	–		

S1.2. CSA variability across individual contrasts

Figure S1 shows the variability in the spinal cord CSA across six contrasts on the test set (*n=49*; n_{vol}*=294*) of the spine-generic multi-subject database (Cohen-Adad et al., 2021a) between three methods: (i) `sct_deepseg_sc` (Gros et al., 2019), (ii) `contrast-agnostic_v2.0` (Bédard et al., 2025), and the proposed model, `contrast-agnostic_v3.0`. Compared to the previous version (`v2.0`), our proposed model (`v3.0`) achieves a similar CSA variability on the test set of healthy participants despite being trained on heterogeneous data containing new contrasts and several pathologies.

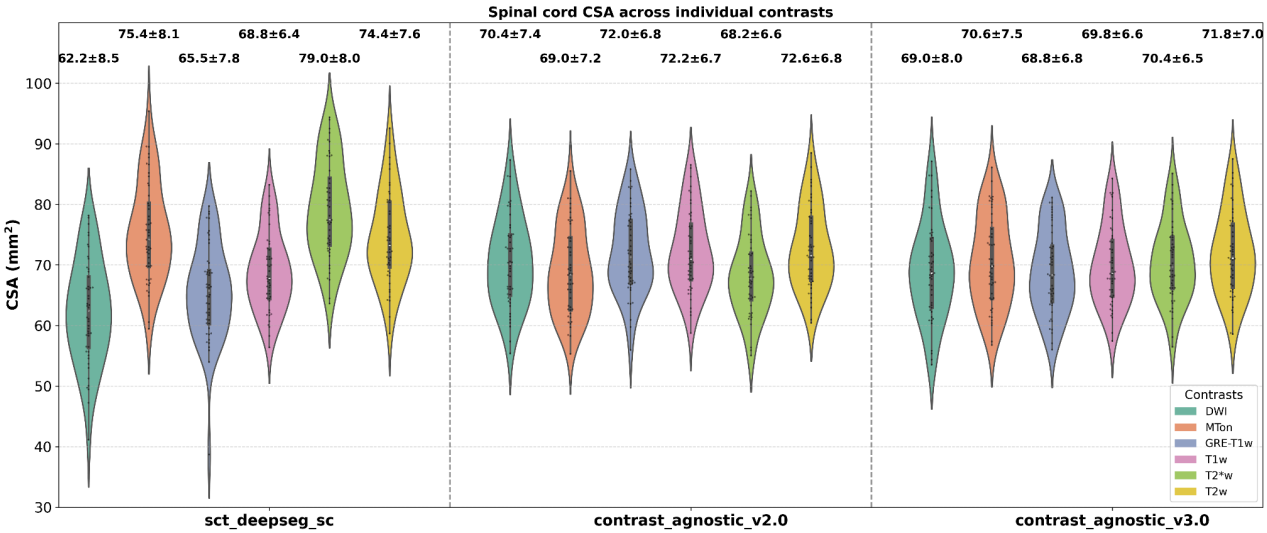

Figure S1. *Variability in spinal cord CSA across 6 contrasts compared with existing automatic segmentation methods on a test set of healthy participants (n=49). Even after the addition of new pathologies and contrasts to the training set, CSA variability achieved by the proposed `contrast-agnostic_v3.0` model remains similar to `contrast-agnostic_v2.0` (trained only on a healthy participants database) and shows a substantial improvement over `sct_deepseg_sc`.*

S1.3. CSA variability across scanner manufacturers

In this section, we evaluate the variability in the CSA measurements for a *single subject* across different scanner manufacturers. We used the spine-generic `data-single-subject` dataset (Cohen-Adad et al., 2021), which includes cervical spinal cord scans in a single healthy participant using six contrasts (T2w, T1w, T2*w, MT-on, GRE-T1w, and DWI) across 15 sites with 3 scanner vendors (GE; *n=4*, Philips; *n=4*, Siemens; *n=7*). As with the previous evaluations, we compared three methods: `sct_deepseg_sc` (Gros et al., 2019), `contrast_agnostic_v2.0` (Bédard et al., 2025), and the proposed `contrast_agnostic_v3.0`, for contrasts and sites. In all comparisons, the spinal cord segmentations were obtained independently for each of the above methods, and the vertebral levels were identified using `sct_label_vertebrae`. Then, we calculated the CSA averaged across C2-C3 vertebral levels and computed its standard deviation (STD) across scanner manufacturers.

It is important to stress that all data points represent the *same* participant. Each of the 6 contrasts compares the two segmentation methods across all 15 sites. **Figure S2** presents the CSA STD across 6 contrasts per site for both segmentation methods, separated per MRI vendor. The STD using the `contrast_agnostic_v3.0` method yields a lower STD than when using `sct_deepseg_sc` for segmentation and is very similar to `contrast_agnostic_v2.0`.

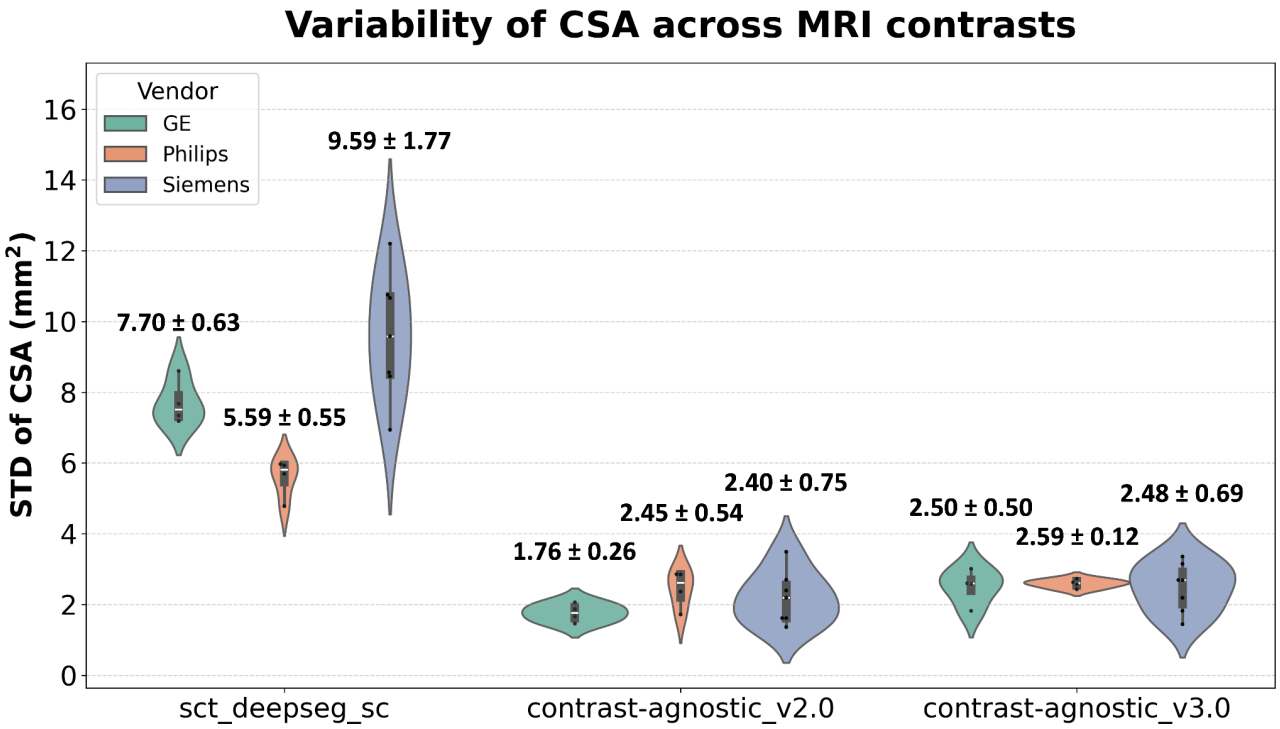

Figure S2. *Variability of spinal cord CSA across contrasts separated per vendor for segmentations generated with sct_deepseg_sc (Gros et al., 2019), contrast-agnostic_v2.0 (Bédard et al., 2025) and contrast-agnostic_v3.0 (proposed) segmentation and contrast-agnostic of the same participant scanned across 15 different MRI sites. Each dot represents one site; mean and standard deviation are presented above.*

S1.4. CSA variability with recursively generated labels

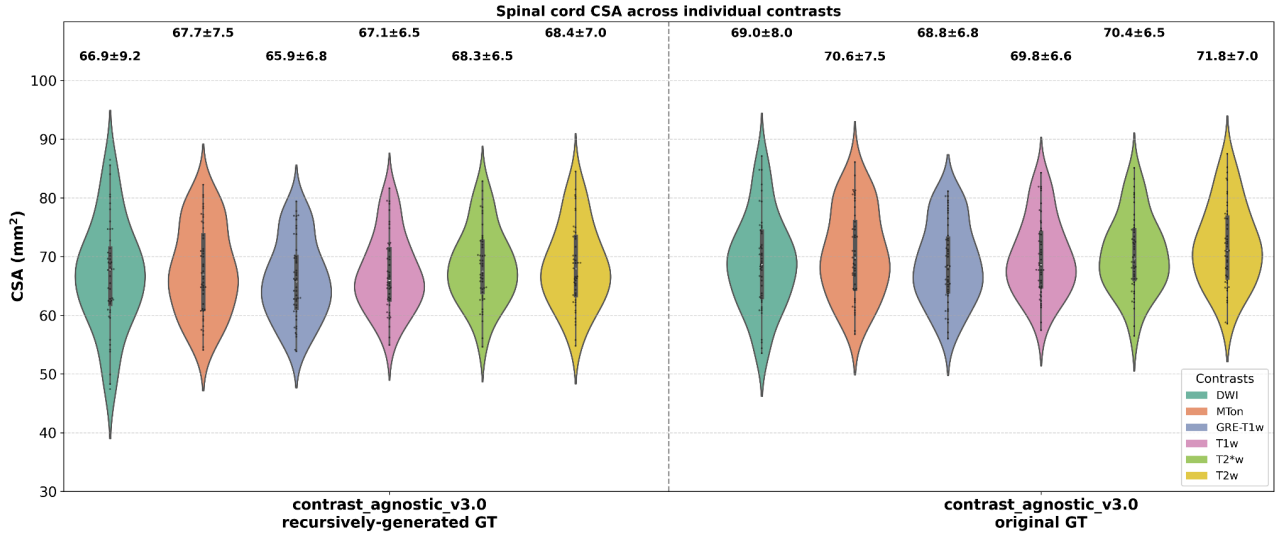

Figure S3. *Variability in spinal cord CSA across 6 contrasts on a test set of healthy participants (n=49) compared between the models trained with the: (i) original distribution of GT masks created*

from a mix of manual annotations and automatic segmentation methods, and (ii) GT masks regenerated with contrast_agnostic_v3.0 model without any manual corrections. The model trained on recursively generated GT masks achieved a lower average CSA per contrast compared to the model trained on the original distribution of GT masks on all contrasts.

Figure S3 plots the average CSA per contrast for the ablation study, comparing the downstream effect of training the contrast_agnostic_v3.0 model on the original distribution of GT masks and the masks generated recursively without any manual correction.