Titre: Title:	Towards Automatic Spinal Cord MRI Analysis for Improved Estimation of Imaging Biomarkers
Auteur: Author:	Muni Venkata Naga Karthik Enamundram
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Enamundram, M. V. N. K. (2025). Towards Automatic Spinal Cord MRI Analysis for Improved Estimation of Imaging Biomarkers [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/70362/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/70362/
Directeurs de recherche: Advisors:	Julien Cohen-Adad, & Sarath Chandar Anbil Parthipan
Programme: Program:	Génie biomédical

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Towards Automatic Spinal Cord MRI Analysis for Improved Estimation of Imaging Biomarkers

MUNI VENKATA NAGA KARTHIK ENAMUNDRAM
Institut de génie biomédical

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie biomédical
Octobre 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

Towards Automatic Spinal Cord MRI Analysis for Improved Estimation of Imaging Biomarkers

présentée par **Muni Venkata Naga Karthik ENAMUNDRAM**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Farida CHERIET, présidente
Julien COHEN-ADAD, membre et directeur de recherche
Sarath Chandar ANBIL PARTHIPAN, membre et codirecteur de recherche
Hervé LOMBAERT, membre
Ismail BEN AYED, membre externe

DEDICATION

To my parents, friends, and loved ones
who stood by me through thick and thin.

"The reward of our work is not what we get, but what we become."

– Paulo Coelho, Manuscript Found in Accra

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, Prof. Julien Cohen-Adad, for his guidance and support throughout my PhD. In June 2021, I cold-emailed Julien about doing a PhD in his lab, and I remember our first meeting discussing what I had worked on during my master's and some ideas for potential PhD projects. He decided to take me on as his PhD student right there during the meeting, and here we are four years later. Never underestimate the power of a cold email, as it can change the trajectory of your life. Julien has been incredibly supportive and was always available (even on short notice!) to discuss research and give academic advice at times when I was stuck. I am also thankful for his patience in correcting my writing. I have received some of the best writing advice from him, especially on getting to the point quickly (which I tend to *not* do), and I will keep this with me for the rest of my life. Apart from research and life in academia, there are other things that you learn implicitly from your supervisor. I highly admire his drive and passion for maintaining open-source software and doing good science, and that has made me strive harder during my PhD and deliver good work. I will fondly remember these four years at NeuroPoly; thank you for everything.

I would also like to thank my co-supervisor, Prof. Sarath Chandar. I remember being extremely anxious about my first meeting with Sarath in October 2021 because of his short two-part interview process consisting of technical and coding interviews. I somehow made it, and am glad I was able to be a part of his lab. Being part of two labs with completely different research foci exposed me to a variety of research topics and ideas. I am grateful to both of my supervisors for providing me with this opportunity during my PhD. Special thanks to Amal Bennani, staff member at Polytechnique, for her support in making the administrative side of PhD easier.

Next, I would like to thank my master's supervisor, Prof. Catherine Laporte. I spent two years at Cathy's lab at ÉTS and fondly reminisce about that time whenever I walk past the ÉTS buildings in downtown Montréal. Before starting my PhD, I had asked her for the definition of a good PhD thesis, and she had said, "A good PhD thesis is a finished PhD thesis." At the time, I was too naïve to understand this statement. As I am writing this acknowledgment, I can fully grasp the depth of her statement, having gone through the ups and downs of a PhD. I also thank my undergraduate thesis supervisor at Shiv Nadar University, Prof. Madan Gopal, with whom I worked on my first deep learning project. Lastly, I would also like to thank Prof. Ajit Rajwade from the CSE department at IIT

Bombay. It was during a brief internship at his lab that I was first exposed to various topics in image processing, and since then I have only been more driven to pursue both master's and PhD degrees in this field.

Apart from your supervisors, your labmates are the next most important set of people during your PhD. I have been quite fortunate in having wonderful labmates (who have also become incredibly close friends). At NeuroPoly, we had a great time during lunches, regular "5 à 7" events, and board game evenings at the lab. I will fondly remember our trips to Université Laval for the QBIN workshop and to Singapore for the ISMRM conference. Thank you Alex, Alexia, Andréanne, Andjela, Armand, Daniel, Emma, Jan, Joshua, Marie-Hélène, Mathieu, Nadia, Nathan, Nilser, Pierre-Louis, Rohan, Samuelle, Sandrine, and Thomas. I also used to spend two days working from Mila and attend lab meetings with Sarath's group. I especially liked the fun events at Chandar lab, where I first discovered some of the winter traditions in Québec like snowtubing, going to a sugar shack, etc. I thank Abdel, Alex, Ali, Arjun, Behnoush, Darshan, David, Doriane, Gonçalo, Hadi, Jana, Lola, Mathieu, Megh, Milan, Prashant, and Quentin for making me feel a part of the lab. My research visit to TUM was one of the best times of my PhD. I got to be a part of two labs even in Munich, and I cherish those moments to this day. Thank you CuiCi, Tun, Julian, Felix, Laura, Niccolo, Markus for making me feel welcomed. I will remember our trips to Fino to get lunch and "peanut butter pasta" Thursdays. Thank you Sophie, Vasiliki, Tamara, Yundi, and Harvey for making me feel like a part of Daniel's lab. I may have missed several others who I have briefly met; I thank all of you for being a part of this journey.

I had great research collaborators during my PhD. Thank you Dario Pfyffer, Lynn Farner, Simon Schading, Anna Lebret, Charidimos Tsagkas, Emanuele Pravatà, Gergely David, Prof. Andrew Smith, Prof. Kenneth Weber, Prof. Patrick Freund, and Prof. Cristina Granziera for all the research discussions we have had. Your interest and excitement towards my projects kept me motivated. I would also like to thank Prof. Mark Mühlau and Prof. Daniel Rueckert for being such wonderful hosts during my stay in Munich.

I can hardly overstate the importance of having a circle of good friends during your PhD. They have kept me sane during difficult times. Thank you, Jan, for your collaboration in our SCI projects. You had a big influence in how I approach data analysis and writing code now. I will miss our table tennis sessions (if people ask, I am going to lie and tell them that I won most of the times). Thank you, Julian, for ensuring that I had memorable stay at Munich, for being there to discuss ideas and give advice on what to focus on, and for our memorable trips to Denmark and Morocco. I will not forget the impulsive decision to do the Læsø ultra half marathon. It was well worth it. Thank you Karan and Akshay for showing me around

Chicago. Thank you, Karan, Ishmeet, Aparna, Vidyank, Akhilesh for making it to the Euro trip. I cannot wait to plan for the next one. Thank you Ayushman for letting me crash at your place in Paris. Thank you, Pranshu, for being a nice roommate. I can safely say that my knowledge of world history has improved thanks to our discussions on this topic.

Without the support of my parents and close relatives, I would have never made it this far. Thank you, Ramya Akka and Srikanth Anna, for showing me around New York and Seattle. You made my first visit to the USA quite memorable. Simply thanking Swathi Didi and Tanmai Bhaiyya would not do justice to how much they have cared for me. I had a lot of fun exploring the province of Ontario with you both – those late nights spent swiping and zooming in and out on the map of Canada, searching for remote Airbnbs, and long drives through small, picturesque Canadian towns. If given another chance, I would do all of that exactly the same. Thank you for being available for our weekly calls; they have played a huge role in keeping me sane.

Lastly, thank you, Nanna and Ammi, for your unending support. This year marks my tenth year away from home. I remember it was in July 2015 that we were packing our suitcases to drop me off at SNU. The passage of time is paradoxical: all these years seem to have gone by like a breeze, yet we know exactly the struggles we've overcome. I cannot comprehend the psychological sacrifices you've made during these last 10 years I was away from home. Thank you for being supportive and understanding. This thesis is for both of you.

I was fortunate to hold scholarships and receive funding from several agencies: Fonds de Recherche du Québec (FRQNT), UNIQUE Québec, Quebec Bio-Imaging Network (QBIN), and Deutscher Akademischer Austauschdienst (DAAD) Short-term Research Grants. I would like to thank all the funding agencies for their financial support throughout my PhD.

RÉSUMÉ

L'analyse automatique des images de la moelle épinière fournit des informations essentielles sur sa morphologie, permettant une caractérisation cohérente des variations anatomiques au sein des populations et facilitant le diagnostic et le suivi des maladies neurologiques. Les mesures morphométriques issues des examens d'imagerie par résonance magnétique (IRM) de la moelle épinière constituent des biomarqueurs clés pour améliorer notre compréhension de la physiopathologie de diverses maladies. La segmentation de la moelle épinière et des lésions est une condition préalable essentielle à l'extraction de biomarqueurs cliniques. Cependant, les défis inhérents à l'imagerie de la moelle épinière, à savoir la petite taille de la structure physique de la moelle, sa sensibilité au bruit et aux artefacts de mouvement, combinés à des effets de volume partiel, rendent la segmentation automatique extrêmement difficile. Par conséquent, les outils existants manquent de robustesse et ne généralisent pas correctement, ce qui suggère la nécessité de disposer d'outils automatiques capables de segmenter la moelle épinière à travers plusieurs contrastes IRM et pathologies.

L'objectif général de cette thèse est de développer des outils automatisés et généralisables pour la segmentation robuste de la moelle épinière et des lésions à travers différents contrastes IRM et pathologies afin de mieux estimer les biomarqueurs d'imagerie. Cet objectif se décline en trois contributions : (i) une méthode spécifique au contraste pour la segmentation des lésions intramédullaires de la moelle épinière, (ii) une approche plus généraliste et indépendante du contraste pour la segmentation de la moelle épinière, et (iii) un cadre conçu pour faciliter l'entrainement continue des modèles de segmentation de la moelle épinière.

Dans le cadre de notre première contribution, nous présentons SCIseg, un outil de segmentation des lésions et de la moelle épinière, qui vise à mesurer automatiquement les biomarqueurs cliniques associés à la récupération fonctionnelle chez les patients souffrant de lésions traumatiques de la moelle épinière. SCIseg a été développé à partir de données recueillies sur trois sites et a utilisé une stratégie d'apprentissage actif avec intervention humaine pour annoter progressivement les ensembles de données destinés à l'entraînement. Les résultats ont montré que les biomarqueurs IRM mesurés à partir des prédictions de SCIseg et des masques annotés manuellement ne présentaient aucune différence significative, prouvant ainsi la fiabilité des prédictions automatiques. Depuis sa sortie, SCIseg a été validé sur des cohortes externes inédites, démontrant ainsi son utilité clinique.

Pour la deuxième contribution, nous avons développé un modèle de segmentation de la moelle épinière indépendant du contraste, en mettant particulièrement l'accent sur la réduction de

la variabilité des mesures morphométriques de la moelle épinière entre différents contrastes. Notre approche a utilisé des masques de segmentation non-binaire pour l'entraînement, obtenus à partir d'un nouveau cadre de prétraitement conçu pour générer des masques probabiliste uniques pour chaque contraste pour un sujet donné. La comparaison avec des méthodes de généralisation de domaine et des modèles spécifiques au contraste a montré l'impact de l'entraînement avec des masques probabiliste sur la réduction de la variabilité morphométrique entre les contrastes. Une segmentation robuste de la moelle épinière est essentielle pour comprendre les mécanismes de diverses maladies neurologiques. Depuis sa sortie, notre outil a été utilisé pour mesurer l'atrophie de la moelle épinière dans des études longitudinales impliquant des patients atteints de myélopathie cervicale dégénérative.

Notre troisième contribution consiste à appliquer des pratiques de développement logiciel pour créer un cadre de formation continue pour les modèles de segmentation de la moelle épinière. Nous présentons un scénario d'apprentissage continu en production, dans lequel (i) nous avons développé et déployé un modèle de segmentation formé à partir de données multi-institutionnelles recueillies sur 75 sites, incluant 9 contrastes IRM, (ii) un workflow GitHub Actions surveille la dérive morphométrique entre les différentes versions du modèle, et (iii) dans le cadre d'une application concrète, nous avons mis à jour une base de données normative de participants en bonne santé contenant des mesures couramment utilisées de la morphométrie de la moelle épinière. À mesure que les ensembles de données médicales évoluent au fil du temps et que les modèles de segmentation continuent d'être développés, il est essentiel de garantir une dérive minimale des performances entre les différentes versions des modèles. Notre approche présente une preuve de concept dans ce sens.

En conclusion, les méthodes présentées dans cette thèse améliorent l'état actuel des modèles de segmentation de la moelle épinière, fournissant des outils robustes pour l'estimation objective des biomarqueurs cliniques et permettant le développement continu de modèles de segmentation à l'avenir.

ABSTRACT

Automatic image analysis of the spinal cord offers key insights into its morphology, enabling consistent characterization of anatomical variations across populations and facilitating the diagnosis and monitoring of neurological diseases. Morphometric measures derived from spinal cord magnetic resonance imaging (MRI) scans serve as key biomarkers in improving our understanding of the pathophysiology of various diseases. Segmentation of the spinal cord and lesions is a crucial prerequisite for extracting clinical biomarkers. However, the inherent challenges associated with spinal cord imaging, namely, the small physical structure of the cord, its susceptibility to noise and motion artifacts combined with partial volume effects make automatic segmentation extremely challenging. Consequently, existing tools lack robustness and are unable to generalize, suggesting the need for automatic tools that can segment the spinal cord across several MRI contrasts and pathologies.

The overarching goal of this thesis is to develop generalizable, automatic tools for the robust segmentation of the spinal cord and lesions across various MRI contrasts and pathologies for better estimation of imaging biomarkers. This goal is spread across three contributions: (i) a contrast-specific method for segmenting intramedullary lesions in spinal cord injury, (ii) a more generalist, contrast-agnostic approach for spinal cord segmentation, and (iii) a framework designed to facilitate continuous training of spinal cord segmentation models.

As the first contribution, we introduce SCIseg, a tool for spinal cord and lesion segmentation in spinal cord injury (SCI), aiming to automatically measure clinical biomarkers that have been associated with functional recovery in patients with SCI. SCIseg was developed using data gathered from three sites and used human-in-the-loop active learning strategy to incrementally annotate the datasets for training. Results showed that the MRI biomarkers measured from SCIseg predictions and manually-annotated masks had no significant difference, proving the reliability of automatic predictions. Since its release, SCIseg has been further validated on unseen, external cohorts demonstrating its clinical utility.

For the second contribution, we developed a contrast-agnostic spinal cord segmentation model, with a particular focus on reducing the variability in morphometric measures of the spinal cord across various contrasts. Our approach used soft segmentation masks for training obtained from a novel preprocessing framework designed to generate unique, soft masks for each contrast for a given subject. Comparison with domain generalization methods and contrast-specific models showed that impact of training with soft masks towards reducing morphometric variability across contrasts. Robust segmentation of the spinal cord holds the

key to understanding various neurological diseases mechanisms. Since its release, our tool has been used to measure the atrophy of the spinal cord in longitudinal studies involving patients with degenerative cervical myelopathy.

Our third contribution applies software development practices to create a continuous training framework for spinal cord segmentation models. We present in a lifelong-learning-in-production scenario, where, (i) we developed and deployed a segmentation model trained on multi-institutional data gathered from 75 sites including 9 MRI contrasts, (ii) a GitHub Actions workflow monitors for morphometric drift between various model versions, and (iii) as a real-world application, we updated a normative database of healthy participants containing commonly used measures of spinal cord morphometry. As medical datasets evolve over time and segmentation models continue to be developed, ensuring minimum drift in performance across various model versions is critical. Our approach presents a proof-of-concept in this direction.

In summary, the methods presented in this thesis improve the current state of spinal cord segmentation models, providing robust tools for objective estimation of clinical biomarkers and enabling the continuous development of segmentation models in the future.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

AL	Active Learning
ASD	Average Surface Distance
CE	Cross Entropy
CNS	Central Nervous System
CNN	Convolutional Neural Networks
CI/CD	Continuous Integration and Continuous Delivery
CL	Continual Learning
CSA	Cross-sectional Area
CSF	Cerebrospinal Fluid
DCM	Degenerative Cervical Myelopathy
DWI	Diffusion-weighted Imaging
DL	Deep Learning
EPI	Echo Planar Imaging
GRE	Gradient Echo
GT	Ground Truth
LL	Lifelong Learning
MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
PPV	Positive Predictive Value
PSIR	Phase-sensitive Inversion Recovery
PVE	Partial Volume Effect
QC	Quality Control
RF	Radiofrequency
RVE	Relative Volume Error
SE	Spin Echo
SCI	Spinal Cord Injury
SCT	Spinal Cord Toolbox
SNR	Signal-to-Noise Ratio
STIR	Short Tau Inversion Recovery
T1w	T1-weighted
T2w	T2-weighted
ViT	Vision Transformer

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

1.1 Medical image segmentation

Image segmentation is a key element in image processing and computer vision defined as the process of dividing the image into semantically meaningful objects or regions of interest [19]. Depending on the context and the application, these regions could be everyday objects like cars, trees or pedestrians in natural images or anatomical organs in medical images. The rise of deep learning (DL) has resulted in a growing proportion of segmentation algorithms employing neural network-based approaches, which are outperforming traditional methods by large margins on popular benchmarks measuring segmentation accuracy and speed [20–22]. Image segmentation as a task can be further classified into two categories: *semantic segmentation* and *instance segmentation*. While both tasks can fundamentally be formulated as a dense pixel-wise classification problems, the outputs of these tasks differ. Semantic segmentation algorithms output pixel-wise labels, assigning an integer value (i.e. a *class*) to each pixel in the image. On the other hand, instance segmentation algorithms go beyond the class-level pixel-wise labels and assign labels at the individual object-level. Unlike image classification where the whole image is assigned a label, image segmentation is a harder task as pixel-wise labels are assigned to different object categories [22].

Image segmentation lies at the heart of several applications including autonomous driving, satellite and aerial imagery, medical image analysis, robotics, and augmented reality [21, 22]. Zooming in on the medical domain, automated medical image segmentation plays a key role in tasks requiring computer-aided diagnosis [23–25]. For instance, in disease monitoring, segmentation can help in tracking the growth/shrinkage of tumors/lesions across longitudinal time points from the patients' medical exams, thus aiding in designing appropriate treatments [26–29]; in surgery, segmentation can help in visualizing complex structures like blood vessels, nerves, and organ boundaries during surgical procedures and provide real-time guidance during minimally invasive surgeries by overlaying segmented image data onto live surgical views, helping surgeons navigate around critical structures [30, 31]. Accurate segmentation can also offer objective and standardized indicators for clinical trials in various diseases.

1.2 Imaging modalities

The performance of the segmentation algorithms is ultimately bound by the image quality, directly influencing diagnostic precision and treatment planning. Commonly-used imaging

modalities include Computed Tomography (CT), Magnetic Resonance Imaging (MRI), X-ray, Ultrasound (US), each with its unique advantages and challenges [32]. CT scans offer excellent bone definition, making them ideal for segmenting skeletal structures, but struggle with soft tissue contrast. MRI provides superior soft tissue differentiation through multiple sequences (*e.g.* T1, T2, etc.), enabling precise segmentation of soft tissue tumors or lesions in the brain or spinal cord, though it suffers from motion artifacts. Ultrasound delivers real-time imaging and avoids radiation exposure, but its inherent speckle noise and operator-dependency complicate segmentation tasks. X-rays, while widely accessible, present significant challenges for segmentation due to their 2D projection nature and superimposed structures. Figure 1.1 shows some examples of segmentation of different organ systems across various imaging modalities.

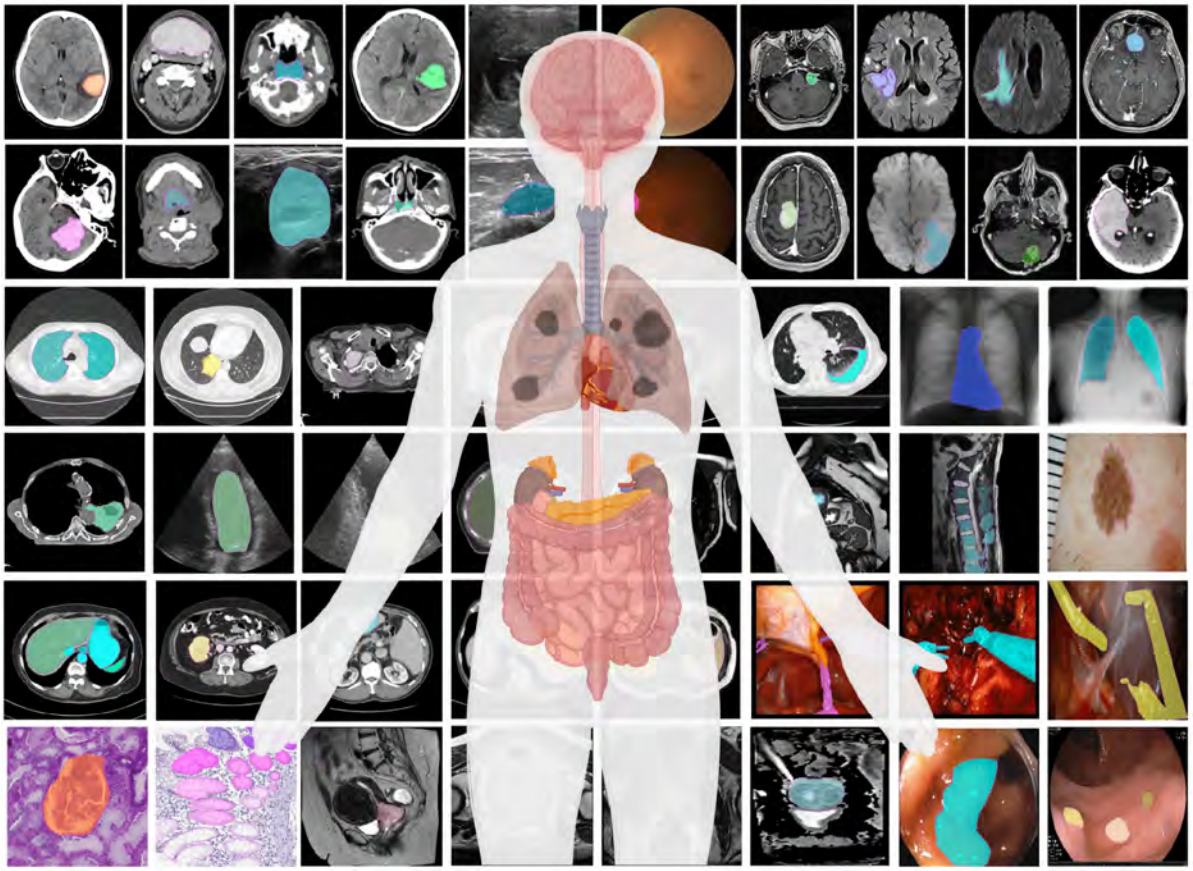

Figure 1.1 Examples of segmentation across different imaging modalities and anatomies. Segmentation masks are overlaid as the regions of interest on the original image. Source: Ma *et al.* [1].

1.3 Magnetic resonance imaging

Out of the common imaging modalities, MRI has become the gold standard in diagnosis and prognosis of diseases of the central nervous system (CNS) affecting the brain and spinal cord [33–36] due to its excellent visualization of soft tissue contrast, multi-planar capability to image directly in any plane, and the absence of ionizing radiation. Different MRI sequences such as T1-weighted, T2-weighted, DWI (Diffusion-Weighted Imaging) and contrast-enhanced scans provide complementary information about tissue structure, water content, inflammation, and blood flow. For example, T2-weighted sequences are highly sensitive to fluid changes and are useful for detecting edema, demyelination, and lesions commonly seen in multiple sclerosis (MS) and spinal cord injury (SCI). Advanced contrasts like PSIR (Phase-Sensitive Inversion Recovery) improve the visualization of cortical and spinal cord lesions, especially in MS, due to its superior gray-white matter contrast [37, 38]. Overall, various sequences complement each other for more accurate diagnoses, especially in diseases with subtle or complex presentations.

While the spinal cord is merely cast as a channel for relaying information between the brain and the body, it is becoming apparent that a thorough characterization of the CNS cannot be achieved without insights into spinal cord anatomy and function [39]. Yet, the spinal cord has been largely overlooked by the neuroimaging community, evident by the wealth of research investigating the brain in neurological disorders [39–42]. While there are multitude of reasons why the progress has been imbalanced, the following section discusses the key challenges in automatic image analysis of the spinal cord.

1.4 Challenges in automatic spinal cord image analysis

Anatomy, Field inhomogeneities, and Physiological noise Imaging of the spinal cord presents with three inherent challenges [39,43,44] (see Figure 1.2): (i) small physical structure of the cord, (ii) magnetic field inhomogeneities, and (iii) susceptibility to noise and motion artifacts. The spinal cord is a long, tiny structure lying within the spinal canal surrounded by the cerebrospinal fluid (CSF), and cartilaginous discs between the vertebral bodies [43, 45]. Its small cross-sectional diameter ($\sim$ 1 cm) requires high spatial resolution (meaning, smaller voxel sizes) to reliably depict anatomical details and minimize partial volume effects, making the acquisitions susceptible to low signal-to-noise (SNR) ratio. Inhomogenous magnetic field around the spinal cord is also one of key challenges. Differences in the magnetic susceptibility of bones, CSF, soft tissues, and air surrounding the spinal cord could result in image artifacts including loss of signal intensity and image distortion [46]. Lastly, the proximity of the

spinal cord to the heart, lungs and other visceral organs that undergo periodic movements is a significant source of noise. The movement of the cord within the spinal canal combined with the pulsatile CSF flow further hamper the MR imaging of the cord. Consequently, these challenges underscore the need for standardized acquisition and robust post-processing techniques to ensure good quality MRI scans for precise delineation of the cord and subsequent quantitative analyses.

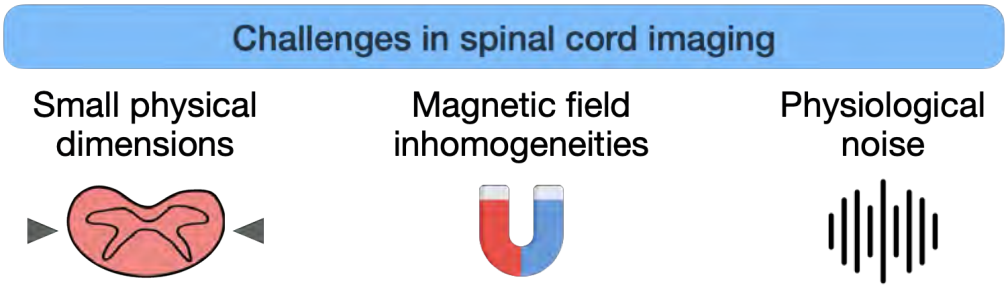

Figure 1.2 Three main challenges in spinal cord imaging.

Partial volume effects Partial volume effect (PVE) is characterized by the mixing of signals from different tissues within the same voxel, resulting in averaged intensities which are not representative of any of the underlying tissues [47–51]. In spinal cord imaging, this phenomenon arises when a voxel is at the interface between CSF and white matter, white matter and gray matter, CSF and vasculature, or white matter and vasculature [44]. Within such voxels, signals originating from tissues with varying spin densities collectively contribute to the overall MR signal, resulting in blurred or indistinct tissue boundaries. PVE can be minimized by increasing the spatial resolution at the cost of lowering the SNR. Imaging using scanners with higher magnetic fields (*e.g.* 7T) can improve spatial resolution [40].

Labeling Manually annotating spinal cord scans is a tedious process. Depending on the image resolution (e.g., axial scans with thick sagittal slices, isotropic, or sagittal scans with thick axial slices) annotating each slice takes considerable amount of time, resulting in a major bottleneck for training segmentation models. Manually-annotated labels suffer from intra-rater and inter-rater variability especially at the cord/lesion boundaries, further complicating the segmentation task [52–55]. This raises questions about what constitutes a *ground truth* (GT) annotation and how the variability in GT annotations limits the performance of automatic methods. Human-in-the-loop deep learning explores concepts that can be used to develop DL systems capable of learning from human feedback [56–58]. Through techniques such as active learning [59, 60] and transfer learning [61], the objective is to choose

the right subset of samples to annotate (either by humans, off-the-shelf pretrained models, or a semi-automatic combination of both) so that segmentation models can be trained on larger datasets. Figure 1.3 illustrates the advantage of human-in-the-loop active learning for aggregating several small unlabeled datasets.

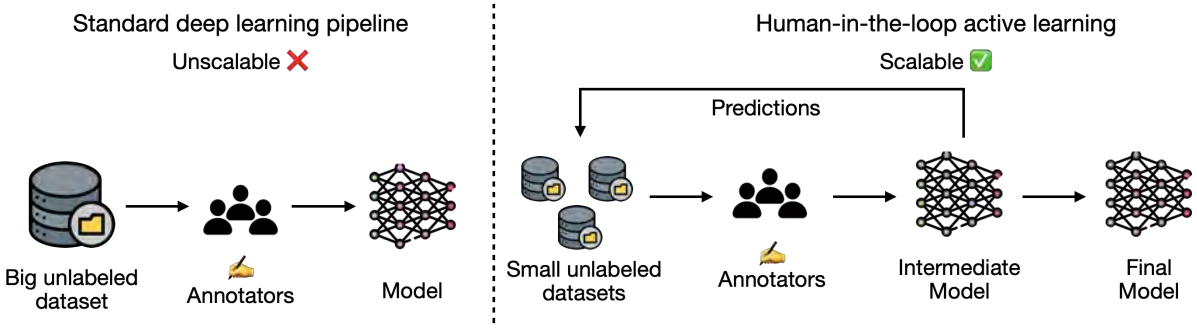

Figure 1.3 Overview of the standard DL training and the human-in-the-loop active learning scenarios. Using predictions from intermediate models makes large-scale DL training more scalable.

Generalization In the context of DL, generalization refers to the ability of a model trained on a particular data distribution (*e.g.*, from a specific scanner or patient population) to perform accurately on unseen data from different sources [62–66]. Complex domain shifts arising from differences in image resolutions, scanners and acquisition protocols prevent current segmentation models of the spinal cord to generalize well across MRI sequences and patient populations. Existing studies have tackled the generalization issue by curating diverse, real-world datasets from multiple sites and via extensive data augmentation [67–69].

Open-source datasets and automatic tools Compared to the wealth of research using brain imaging data, progress in spinal cord imaging is lagging behind partly due to relative lack of open-source datasets [70–74] and automatic tools [75–78] for spinal cord image analysis. Open source challenges such as Brain Tumor Segmentation (BraTS), multiple sclerosis (MS) lesion segmentation [26, 79], held in conjunction with major conferences provide standardized benchmarking datasets for competing teams, ultimately driving the innovation in segmentation algorithms. Furthermore, longitudinal, multi-site studies targeted at specific pathologies such as the Alzheimer's Disease Neuroimaging Initiative [80], Parkinson's Progression Markers Initiative [81], Korean Brain Aging Study [82], help in discovering novel imaging or clinical biomarkers, ultimately improving the outcome of clinical trials and augmenting our understanding of these complex neurological disorders. In addition to the

challenges and datasets, the existence of tools such as FreeSurfer [77], FSL [78], and AFNI [75] facilitates the processing of such large-scale, multi-site datasets.

In contrast, we are now witnessing the gradual shift in the medical imaging community's interest towards the spinal cord, thanks to the standardization of acquisition protocols and few open-source initiatives (*e.g.* spine-generic [42]) and challenges such as the spinal cord gray matter segmentation challenge [83] and the MS multi spine lesion detection challenge[1] [84] and benchmarks such as SPIDER [85] targeted at lumbar spine segmentation. Unlike the plethora of tools available for brain image processing, there do not exist several tools for spinal cord image analysis. Spinal Cord Toolbox (SCT) [86] and JIM[2] are two popular software packages.

Lifelong learning The conventional definition of lifelong learning emphasizes learning new tasks under a continuous stream of input data with the assumption that access to data from earlier tasks is restricted with time [87, 88]. Continual learning is an active area of research with several sub-fields (*e.g.* incremental learning, online learning) proposing methods to tackle various issues (*e.g.* catastrophic forgetting [89, 90]) preventing current DL models to learn *continuously*. Now, consider an alternative view, where the focus is *not* on learning new tasks in a simulated setting, rather, the focus is on the life-cycle of DL models deployed to address real-world tasks. A few key questions arise:

1. As it is typical for medical datasets to grow over time, how can we continuously update the models with new data, once they have been deployed?,
2. As models are trained on different timestamped versions of the datasets, how can we ensure that the drift in models' performance is minimum?,
3. How can we leverage the advancements in software development and machine learning operations (MLOps) to design segmentation frameworks that are equipped to handle continuous streams of input data and can automatically monitor and quantify performance drifts?

In the context of spinal cord segmentation, current models have been developed in static and isolated environments on datasets spanning few MRI contrasts and pathologies. This led to the existence of a collection of models, each specializing in a narrow domain, making it difficult to maintain such models. With the rise of generalist or *foundation* models for

[1] the rationale of the challenge reads: "Lesion dectection: Let's not forget the spinal cord"
[2] https://xinapse.com/jim-9-software/

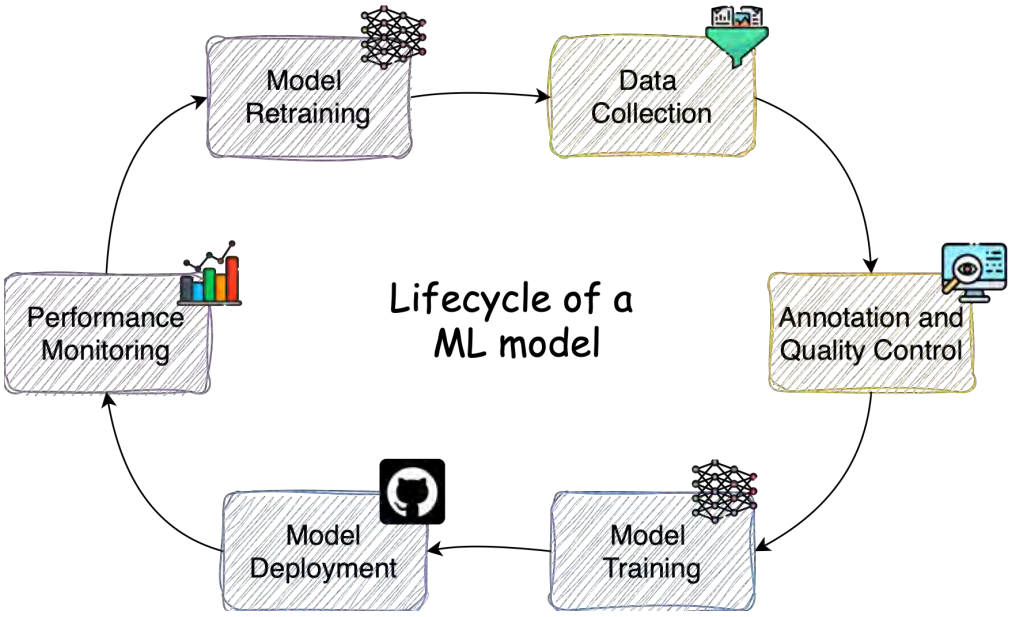

Figure 1.4 Lifecycle of a typical machine learning model in production.

segmentation [1, 66, 91], there is a critical need for designing a standardized segmentation framework facilitating continuous training of a generalist model capable of segmenting the spinal cord across a wide range of contrasts and pathologies, and more importantly, actively monitoring the drift in the performance as the model is updated over time.

Over-reliance on segmentation metrics Progress in the field of automatic medical image segmentation has relied upon the existence of metrics quantifying the similarity between model predictions and GT annotations. In the past three decades, several evaluation metrics were proposed in the literature, however, only a handful of these metrics are currently used [92–94]. Recent studies have shown that lack of thorough validation and the improper choice of evaluation metrics could be one of the major reasons for automatic methods failing to perform outside research environments and be translated into clinical practice [14, 94]. Even if the right metrics are chosen, their mathematical properties are neglected. For instance, the most common overlap-based metric, Dice score coefficient [95], is used evaluate the segmentation of tiny structures (*e.g.* lesions spanning a handful of voxels), when more accurate metrics quantifying lesion-wise statistics exist. There is an over-reliance on the Dice score metric, whether it has clinical value or not. Then, as it is common to have multiple scans per patient, how these data are split into training/testing sets and how they are aggregated to report test scores has the potential to significantly bias the results. The field also suffers from a recency bias towards novel architectures (*e.g.* transformers [96] or state-

space models [97] and ranks architectural advancements (despite them resulting in marginal improvements) over comprehensively validated studies with existing time-tested models such as CNNs [98].

One possible solution to this issue is to tie the segmentation task with a downstream clinical application. For instance, considering the CNS, MRI-derived quantitative biomarkers have long been shown to improve diagnosis of neurodegenerative diseases and structural pathologies of the spinal cord [34, 99–103]. For instance, spinal cord segmentation can be used to compute the cross-sectional area of the cord to quantify atrophy in patients with MS [102,104,105], monitoring cord compression [106], etc. On the other hand, lesion segmentations are used to track and quantify longitudinal changes to the lesion volume [28,107,108], compute width of the spared tissues in traumatic spinal cord injury [101, 109], etc.

1.5 Outline of the thesis

Following a brief introduction to the topics of medical image segmentation and automatic spinal cord image analysis, Chapter 2 expands on each of these topics in greater detail, describing the spinal cord anatomy and defining the fundamental concepts of MRI, image segmentation, active learning and lifelong learning. Chapter 3 establishes the link between the topics introduced in Chapters 1 and 2 and how they fit within research objectives of this thesis. Chapters 4 to 6 present the studies achieving the three research objectives. Chapter 7 provides some perspectives by bringing together the insights from all contributions, presenting the current state segmentation methods, the prospect of application and clinical impact of the work, along with possible avenues for future research. Chapter 8 concludes this thesis.

1.6 Summary of Contributions

This thesis follows the format of a manuscript-based thesis. The list of contributions are divided into two categories: (i) publications that are included in this thesis as individual chapters (Section 1.6.1), and (ii) other contributions from the projects I have participated in during my PhD (Section 1.6.2).

1.6.1 Contributions included in the thesis

Journal Articles / Preprints

- **Naga Karthik, E.***, Valošek, J.*, Smith, A. C., Pfyffer, D., Schading-Sassenhausen, S., Farner, L., Weber II, K.A., Freund, P., & Cohen-Adad, J. (2024). SCIseg: Auto-

matic segmentation of intramedullary lesions in spinal cord injury on T2-weighted MRI scans. Radiology: Artificial Intelligence, 7(1). https://pubs.rsna.org/doi/full/10.1148/ryai.240005

- Bédard, S*., **Karthik, E.N.***, Tsagkas, C., Pravatà, E., Granziera, C., Smith, A., Weber II, K.A. and Cohen-Adad, J., 2025. Towards contrast-agnostic soft segmentation of the spinal cord. Medical Image Analysis, p.103473.. https://www.sciencedirect.com/science/article/pii/S1361841525000210
- **Karthik, E. N.**, Bédard, S., Valošek, J., Aigner, C. S., Bannier, E., Bednařík, J., ... & Cohen-Adad, J. (2025). Monitoring morphometric drift in lifelong learning segmentation of the spinal cord. https://arxiv.org/abs/2505.01364
  (Under review at Imaging Neuroscience; Also presented as a poster at the Conference on Lifelong Learning Agents (CoLLAs) 2025 Workshop paper track)

Conference papers

- **Karthik, E. N.***, **Valošek, J.***, Farner, L., Pfyffer, D., Schading-Sassenhausen, S., Lebret, A., Gergely, D., Smith, A.C., Weber II, K.A., Seif, M., RHSCIR Network Imaging Group, Freund, P.**, & Cohen-Adad, J.** (2024, October). SCIsegV2: a universal tool for segmentation of intramedullary lesions in spinal cord injury. In MICCAI: International Workshop on Applications of Medical AI (pp. 198-209). Cham: Springer Nature Switzerland. ***Poster***
- **Karthik, E. N.**, Bedard, S., Valosek, J., Chandar, S., & Cohen-Adad, J. (2024). Contrast-agnostic spinal cord segmentation: a comparative study of ConvNets and vision transformers. In Medical Imaging with Deep Learning (MIDL) Short-paper Track. ***Poster***
- **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?. In Medical Imaging Meets NeurIPS Workshop. ***Poster***

1.6.2 Other contributions

Journal Articles / Preprints

- **Karthik, E.N.***, McGinnis, J.*, Wurm, R., Ruehling, S., Graf, R., Valosek, J., Benveniste, P.L., Lauerer, M., Talbott, J., Bakshi, R. and Tauhid, S., 2025. Automatic

segmentation of spinal cord lesions in MS: A robust tool for axial T2-weighted MRI scans. Imaging Neuroscience, 3, pp.IMAG-a.
https://doi.org/10.1162/IMAG.a.45

- Lemay, A., Gros, C., **Karthik, E.N.**, & Cohen-Adad, J. (2022). Label fusion and training methods for reliable representation of inter-rater uncertainty. Journal of Machine Learning for Biomedical Imaging (MELBA). 1:1-27.
https://www.melba-journal.org/pdf/2022:031.pdf
- Macar, U.*, **Karthik, E.N.***, Gros, C., Lemay, A., & Cohen-Adad, J. (2021). Team NeuroPoly: Description of the Pipelines for the MICCAI 2021 MS New Lesions Segmentation Challenge. arXiv arXiv:abs/2109.05409.
https://arxiv.org/pdf/2109.05409

Conference Abstracts

- McGinnis J, **Karthik EN**, Rühling SJ, Wurm R, Stern K, Bédard S, Wiltgen T, Zimmer C, Hemmer B, Wiestler B, Rückert D, Kirschke JS, Cohen-Adad J, Mühlau M. (2023). Towards generalizable spinal cord lesion segmentation in multiple sclerosis. 39th Congress of the European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS), Milan, Italy. ***Poster***
- Valošek J, **Karthik EN**, Bouthillier M, Schading-Sassenhausen S, Farner L, Pfyffer D, Smith AC, Weber II KA, Freund P, Cohen-Adad J. (2024). Automatic segmentation of T2-weighted hyperintense lesions in spinal cord injury. Proceedings of the 32th Annual Meeting of ISMRM, Singapore. ***Oral***
- Bédard S, **Karthik EN**, Kaptan M, Kinany N, Van De Ville D, Freund P, Hupp M, Eunyoung Lee L, Traboulsee A, Tam R, Prat A, Kolind S, Oh J, Aigner C, Cohen-Adad J. (2024) Automatic spinal cord segmentation: Generalization across MR parameters, sites, vendors and pathologies. Proceedings of the 32th Annual Meeting of ISMRM, Singapore. ***Oral***
- Bouthillier M, Valošek J, **Karthik EN**, Guay-Paquet M, Guenther N, Rotem-Kohavi N, Humphreys S, Christie S, Fehlings M, Kwon BK, Mac-Thiong JM, Phan P, Cadotte D, Cohen-Adad J. (2024). Building of a National Canadian MRI Repository for Deep Learning Segmentation of Edema and Hemorrhage. Annual Meeting of the Canadian Spine Society, Whistler, Canada.
- McGinnis J, Lauerer M, Wurm R, Graf R, Möller H, **Karthik EN**, Wiltgen T, Berthele A, Zimmer C, Hemmer B, Rückert D, Cohen-Adad J, Wiestler B, Mühlau M, Kirschke

J, Rühling S. (2024). Longitudinal analysis of spinal cord lesion patterns in multiple sclerosis. 40th Congress of the European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS), Denmark. ***Poster***

- Bédard S, **Karthik EN**, Valošek J, Cohen-Adad J. (2025). Minimum sample size to detect spinal cord atrophy with automatic soft segmentation. Proceedings of the 33th Annual Meeting of ISMRM, Honolulu, USA. ***Digital Poster***
- Valošek J, Pfyffer D, **Karthik EN**, Farner L, Schading-Sassenhausen S, Freund P, Cohen-Adad J. (2025). Automatic morphometry of spinal cord injury lesions. 33th Annual Meeting of ISMRM, Honolulu, USA. ***Digital Poster***

CHAPTER 2 BACKGROUND AND LITERATURE REVIEW

In the previous chapter, we briefly introduced the various challenges involved in automatic image analysis of the spinal cord. In this chapter, we will revisit the various topics mentioned within these challenges with the aim of providing sufficient background in understanding the contributions of this thesis. We start with a basic introduction to the spinal cord structure and anatomy. We will then learn the basic principles of MRI physics to understand how MRI machines work and how various MRI sequences are utilized. Using qualitative examples, we will see how various spinal cord pathologies appear in these MRI scans. We will then move on to segmentation and cover the mathematical necessities for understanding the inner workings of modern DL models, key elements of an automatic segmentation framework, followed by an introduction to the concepts of active learning and lifelong learning. By the end of this chapter, the reader should be equipped with the basic knowledge of all the concepts presents in the subsequent chapters of this thesis.

2.1 Spinal Cord

The brain and the spinal cord together constitute the central nervous system. The spinal cord is a long tubular-shaped structure residing in the spinal column, surrounded by cerebrospinal fluid (CSF). It extends from the bottom of the brainstem (at the area called medulla oblongata) to the first or second lumbar vertebrae in the lower back, tapering at the end to form a cone called the conus medullaris (Figure 2.1). As a continuation of the brainstem, the spinal cord is responsible for transmitting nerve signals between the brain and the peripheral nervous system, ensuring the transfer of efferent and afferent messages between the cerebral cortex and the motor and sensory systems. It is also responsible for operating and coordinating reflex actions independent of the brain, for example, controlling rhythmic movements such as breathing or walking.

2.1.1 Anatomy and structure

The spinal column, which houses the spinal cord, is made up of 33 bones called vertebrae. Five vertebrae are fused together to form the sacrum (part of the pelvis), and four small vertebrae are fused together to form the coccyx (tailbone). The spine itself is divided into four regions: (i) cervical vertebrae (C1-C7); located in the neck, (ii) thoracic vertebrae (T1-T12); located in the upper back and attached to the ribcage, (iii) lumbar vertebrae (L1-L5);

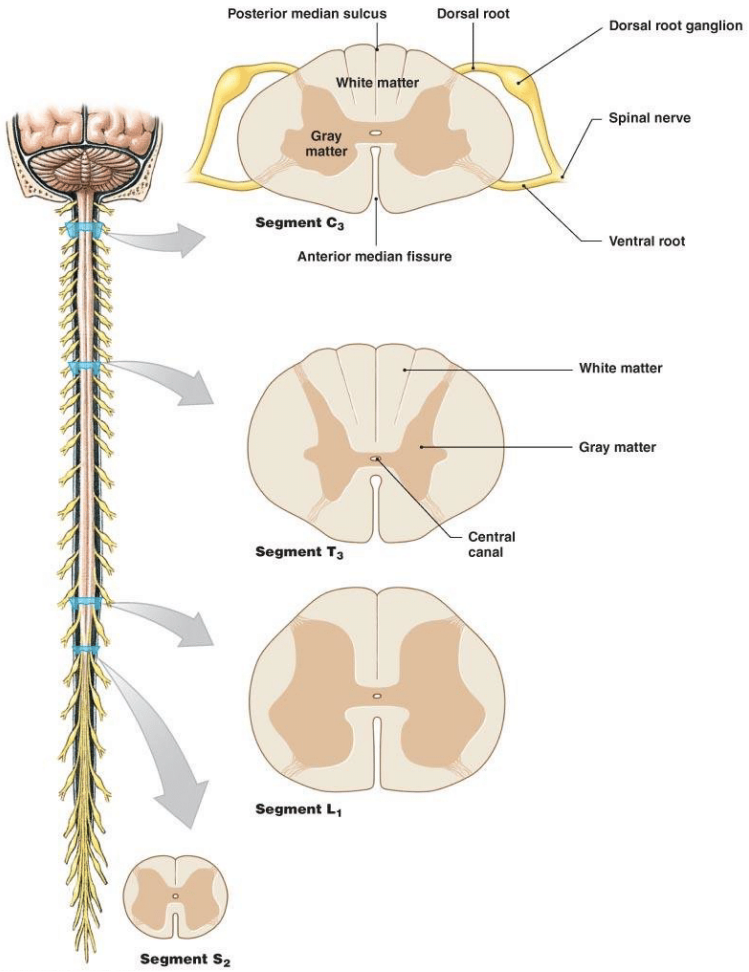

Figure 2.1 Anatomy of the human spinal cord. Source: Martini *et al.* [2].

located in the lower back, and (iv) sacral vertebrae (S1-S5); located in the pelvis. Between the vertebral bodies (except for C1 and C2) are the discs serving as a supportive structure for the spine.

The four regions of the spine are are divided into 31 segments with 31 pairs of spinal nerves. One member of the pair exits on the right side and the other exits on the left. There are 8 cervical nerves, 12 thoracic nerves, 5 lumbar nerves, 5 sacral nerves, and 1 coccygeal nerve. Each nerve exits the vertebral column, passing through the intervertebral foramina to its designated location in the body.

The spinal cord is protected by three layers of tissue or membranes called meninges, that surround the canal. The dura mater is the outermost layer, forming a tough protective coating. Between the dura mater and the surrounding bone of the vertebrae is a space called the epidural space. The arachnoid mater is the middle protective layer. The space between

the arachnoid and the underlying pia mater is called the subarachnoid space/cavity (blue in Figure 2.2). The subarachnoid space is filled with CSF, a clear, colorless fluid that surrounds the cord. CSF plays several key roles [110]: (i) it protects the brain and the spinal cord by providing a fluid buffer that acts as a shock absorber (or a cushion) from mechanical forces, (ii) it helps keeps the brain afloat in neutral buoyancy without being impaired by its own weight, (iii) it prevents brain ischemia and regulates cerebral blood flow, and (iv) it allows for the removal of metabolic waste from the brain. The CSF flows unidirectionally (rostral to caudal) in the ventricular system until it reaches the subarachnoid space, where it becomes multidirectional. The movement of CSF is pulsatile and is driven by the cardiac cycle.

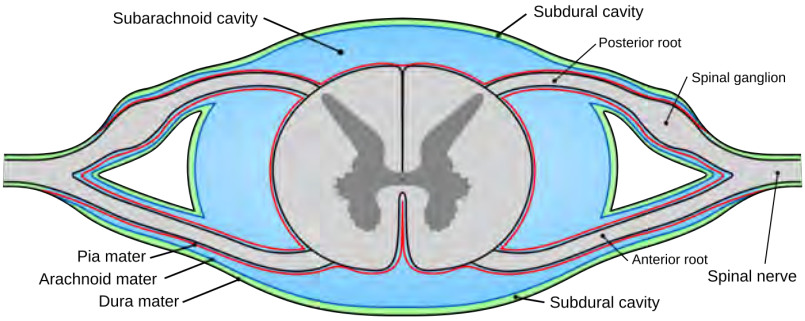

Figure 2.2 Cross-section of the spinal cord with the three layers of tissue that surround the cord. Source: Wikimedia Commons [3].

Grey matter The spinal cord is made up of grey and white matter, appearing as butterfly- or H-shaped grey matter surrounded by white matter. The grey matter is a collection of cell bodies of motor and sensory neurons, interneurons and glial cells. The supports of the "H" make up the dorsal (posterior) and ventral (anterior) horns on the left and right sides (Figure 2.3). Through the center of the spinal cord, running longitudinally, is the central canal filled with CSF. The dorsal horns receive sensory information from afferent nerve fibers that carry sensory signals from the body. These afferent fibers transmit information about touch, pressure, pain, temperature, and other sensations to the dorsal horn. The ventral horns house efferent motor neurons, and their axons, also called efferent nerve fibers, carry signals away from the CNS to peripheral muscles and glands. Specifically, these efferent fibers, originating in the ventral horn, exit the spinal cord through the ventral roots and travel through spinal nerves to reach their target tissues. Afferent and efferent fibers are connected through interneurons in the thin strip of the grey commissure around the central canal.

White matter White matter is composed of interconnecting fiber tracts, which are primarily the myelinated sensory and motor axons. The myelin coating of these axons is a fatty substance that insulates them and speeds up nerve signal transmission. The white matter is organized into tracts. Ascending tracts carry information from the sensory receptors to higher levels of the CNS, while descending tracts carry information from the CNS to the periphery. It is divided into three major columns [111]: (i) dorsal columns, carrying ascending sensory information from somatic mechanoreceptors, (ii) lateral columns, including axons that travel from the cerebral cortex to contact spinal motor neurons (these pathways are also referred to as the cortico-spinal tracts), and (iii) ventral columns, carrying both ascending information about pain and temperature, and descending motor information.

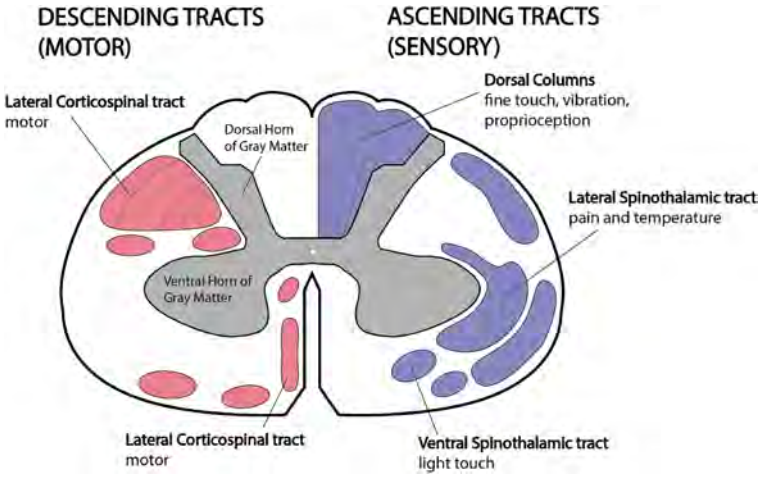

Figure 2.3 Atlas of the major white matter tracts along with dorsal and ventral horns of the grey matter. Source: Grayev [4].

Shape The width of the spinal cord is not uniform across the four regions. To accommodate a greater number of nerve cells and connections to process sensory and motor signals from the upper and lower limbs, the spinal cord is enlarged at the cervical region (between C4-C5 to C7-T1 levels) and lumbar region (between T11 to L2) (Figure 2.4). These enlargements give a characteristic trend to the cord cross-sectional area as shall be seen later in Chapter 6.

2.2 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) uses strong magnetic field and radio waves to create detailed three-dimensional images of the body's internal structures. Unlike CT, MRI does not use any ionizing radiation and also provides better contrast in the images of soft tissues

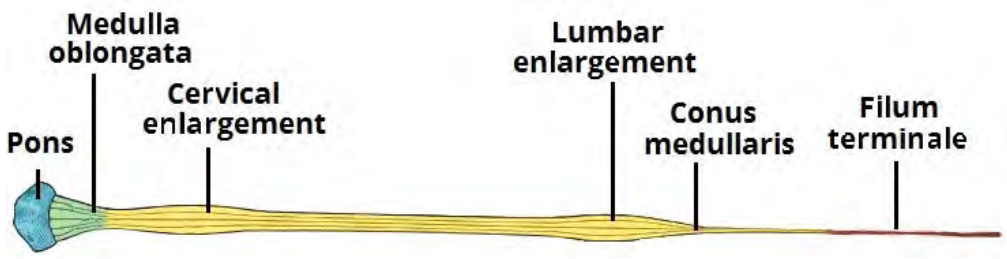

Figure 2.4 Enlargements of the spinal cord. Source: Bath [5]

(*e.g.* brain, spinal cord, muscles, ligaments, and tendons). The core principle behind MRI is nuclear magnetic resonance (NMR), a phenomenon where certain atomic nuclei (*e.g.* like hydrogen atoms in water) absorb and re-emit radiofrequency (RF) energy when placed in a strong magnetic field. By detecting these signals and how they vary across tissues, MRI creates detailed images of the body's structures.

2.2.1 How MRI works

Hydrogen atoms are abundant in humans, particularly in the form of water and fat. The hydrogen nuclei, which consist solely of a proton, are able to absorb and emit RF energy when an external magnetic field is applied. Images are formed by capturing the density of these nuclei in a specific region, which translates to mapping the location of fat and water in the body. As the protons are influenced by fields from atoms to which they are bonded in other tissues, different contrasts between tissues can be obtained based on the relaxation properties of the hydrogen atoms.

During an MRI exam, the participant lies supine on the scanner table and is placed inside the MRI scanner. When a strong magnetic field (B_0) is applied, the protons are aligned to be parallel or anti-parallel to the direction of the field, with a slight majority aligning parallel to B_0 as it is the low energy state. An RF pulse is then passed through the participant, exciting the protons to be out of alignment with the main magnetic field, spinning them out of equilibrium. When the RF pulse is turned off, the protons undergo a rotating motion (*precession*, or, a spiraling decay) returning to their original alignment parallel to B_0. The energy released during the process of realignment is captured by the receiver coils in the scanner. The time it takes for the protons to realign with the magnetic field, as well as the amount of energy released, gives rise to different contrasts in various tissue types. By applying additional magnetic fields (gradients) that vary linearly over space, specific slices to be imaged can be selected, and an image is obtained by taking the 2-D Fourier transform of the spatial frequencies of the signal (k-space). Figure 2.5 illustrates the working mechanism

of MRI.

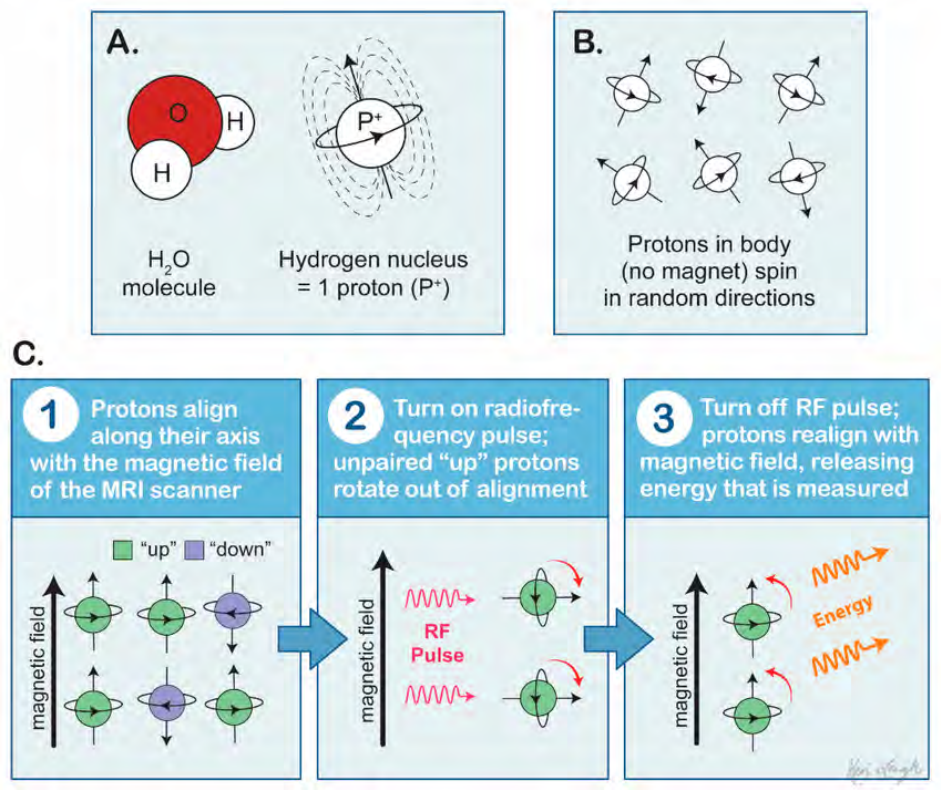

Figure 2.5 Mechanism of MRI. (A) Hydrogen atoms are dispersed within the participant's tissues with intrinsic spin. (B) Hydrogen atoms are spinning in random directions without any alignment. (C) Protons align with the magnetic field in parallel fashion; after the application of a radiofrequency pulse, the protons realign with the magnetic field, releasing energy and generating a high-resolution image of the tissue. Source: Fordham *et al.* [6]

2.2.2 Components of an MRI scanner

The main components of an MRI scanner are [112]: (i) the magnet formed by superconducting coils, (ii) gradient and shim coils, (iii) RF coils, and (iv) computer systems.

Magnet The magnet is the most important and expensive component of the scanner with the rest of the components built around it. The B_0 magnetic field generated inside the scanner is measured in teslas (T). Magnets in most clinical MRI scanners have a field strength of $0.1 - 3$T, with the exception of ultra high field scanners (7, 9.4, 10.5, and 11.7T) used for research purposes [113]. For reference, the Earth's magnetic field is about $25 - 65\mu$T. The core of the magnet is made of a superconducting wire by winding it into a coil and cooling it down to extremely low temperatures (~ 4 Kelvin) with liquid helium. When cooled to such

low temperatures, the wires become superconductive with zero resistance to electric current, achieving a stable magnetic field. The strength of B_0 is an important factor in determining image quality. Higher magnetic fields increase signal-to-noise ratio (SNR), permitting higher resolution or faster scanning. However, higher field strengths require more costly magnets with higher maintenance costs, and have increased safety concerns.

Gradient coils The primary function of gradient coils is to produce additional magnetic fields (called *gradient fields*) that allow spatial encoding of the MR signal. There are three different gradient coils producing three different and less powerful magnetic fields with the purpose of varying the strength of the main magnetic field. Reconstructing an image from MR signals requires proper slice selection and determining the voxels to be designated within the slice. The gradient coils create three sets of magnetic field gradients (x, y, z). When doing 2D imaging, the field gradient in the z-axis (along the direction of B_0) is the slice-selection gradient, while the y-axis and x-axis gradients produce the phase and frequency encodings within the slice, respectively. If the applied RF pulse contains a narrow range of frequencies, then only the region of tissue whose resonant frequencies match those within the RF pulse are excited and go on to contribute to the final image. It is the range of frequencies coupled with the variation induced by the slice-selection gradient that determines the spatial location and thickness of the slab to be imaged.

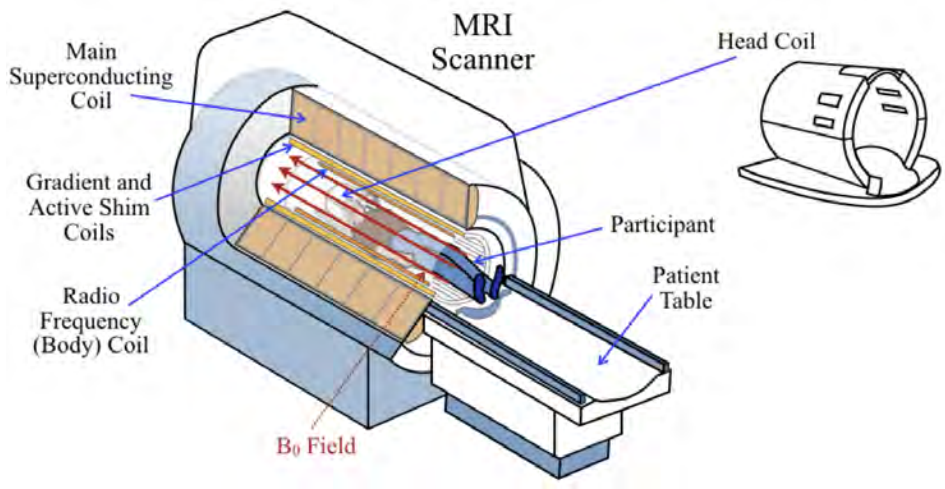

Figure 2.6 Schematic of an MRI scanner, showing main coils and the B_0 field relative to the participant inside the scanner. The head coil (top right) is placed around the participant's head prior to them being moved into the center of the scanner (the bore). Source: Jenkinson and Chappell [7].

Shim coils The main B_0 field created by the superconducting coil should ideally be perfectly uniform, but in practice, has very small inhomogeneities due to the different magnetic properties (*i.e. susceptibility*) of the body's structures. Inhomogeneity in the B_0 field means that the hydrogen nuclei at a given location in a tissue will have a different resonant frequency to what was expected, leading to artifacts such as distortions in the final image. To mitigate the effect of B_0 inhomogeneity, shim coils are built into the scanner to nullify the bulk of inhomogeneities. Since shim coils produce low-rank fields (up to 3rd order spherical harmonics at 7T), they do not fully eliminate the inhomogeneities created by structures with different magnetic susceptibilities.

RF coils Measuring a signal from a collection hydrogen nuclei is only possible when their net magnetization (*i.e.* the sum of the magnetization of all nuclei) is not aligned parallel with the B_0 field. RF coils generate a temporary, oscillating magnetic field, called the B_1 field, by passing a current tuned to the Larmor frequency (the natural wobble frequency of hydrogen nuclei). This B_1 field, oriented perpendicular to the main static B_0 field, effectively "excites" the hydrogen nuclei by gently pushing their magnetization away from the B_0 direction and into the transverse plane. This process, known as *excitation*, is possible because the B_1 field resonates with the nuclei, allowing a much weaker field to achieve this change compared to the strong B_0 field. These precisely timed applications of the B_1 field are called RF pulses and are crucial for obtaining an MRI signal.

Computer system The computer system performs three tasks: (i) controls the main magnet to maintain a uniform B_0 field, (ii) fine-tunes the shim coils to correct for any field inhomogeneities, and (iii) manages the rapidly switching currents in the gradient coils to spatially encode the MRI signal. This coordination ensures that the RF pulses are applied and the resulting signals from the excited nuclei containing their spatial location encoded in their frequencies and phases are recorded in the *k-space*. The raw data in the frequency domain are then converted into the images using the Fourier transform.

2.2.3 MRI sequences

With the primary components of an MRI scanner and their respective functions briefly outlined, the next step is to understand how these elements are coordinated to acquire MRI scans. This is achieved through MRI pulse sequences (or, MRI sequences), which define the temporal application of RF pulses and magnetic field gradients to control the net magnetization and subsequently acquire distinct images. The choice and design of these sequences

are fundamental to determining image contrast, resolution, and sensitivity to various tissue properties. Commonly used classes of MRI sequences include spin-echo, gradient-echo, inversion-recovery, and diffusion-weighted imaging. Each sequence manipulates the longitudinal (T1) and transverse (T2 and T2*) relaxation properties of tissues in unique ways to produce diagnostic images.

Spin echo (SE) SE sequences are designed to compensate for B_0 inhomogeneities and are primarily used to produce images with well-defined T1 or T2 contrast. The classic SE sequence consists of a 90° RF pulse followed, after a delay time TE/2, by a 180° refocusing pulse. This 180° pulse rephases dephased spins, thereby forming a "spin echo" at time TE (echo time). By varying the repetition time (TR) and TE, SE sequences can be tailored to produce T1-weighted (short TR, short TE) or T2-weighted (long TR, long TE) images. T1-weighted images generally make fat appear bright and water dark, while T2-weighted images are highly sensitive to pathology, making water and edema appear bright, as seen in lesions or plaques.

Gradient echo (GRE) GRE sequences differ from SE sequences primarily by using a magnetic field gradient, rather than a 180° RF pulse, to rephase the spins and generate an echo. This omission of the 180° pulse results in faster acquisition times due to shorter TRs. However, GRE sequences are inherently sensitive to T2* relaxation, meaning they are more susceptible to signal loss from B_0 inhomogeneities, such as those caused by air-tissue interfaces or metallic implants. This inherent sensitivity to T2* decay makes GRE sequences particularly useful in visualizing structures that cause local field inhomogeneities, such as microhemorrhages, and calcifications, often referred to as T2*-weighted imaging.

Diffusion-weighted imaging (DWI) DWI is a functional MRI technique that exploits the Brownian motion (diffusion) of water molecules within tissues to generate image contrast. By applying strong, pulsed magnetic field gradients before and after the 180° refocusing pulse (in a modified SE sequence), DWI measures the attenuation of the MR signal caused by the movement of water molecules. In tissues where water diffusion is restricted (*e.g.*, areas of inflammation), the signal attenuation is less, resulting in higher signal intensity on DWI images. In the spinal cord, DWI is technically challenging due to susceptibility to motion, B_0 inhomogeneity, and low SNR. It is typically coupled with Echo Planar Imaging (EPI), a fast image acquisition sequence capable of acquiring an entire images from a single RF pulse. This speed is achieved by rapidly switching the magnetic field gradients to generate a series of echoes that fill k-space in a single shot or a few shots. While highly efficient, EPI is

inherently more sensitive to magnetic susceptibility artifacts and can suffer from geometric distortions due to the existing inhomogeneities.

Inversion recovery Inversion Recovery (IR) sequences are distinguished by an initial 180° RF inversion pulse applied prior to a conventional SE or GRE readout. This pulse inverts the longitudinal magnetization, which then recovers towards equilibrium. The time between the 180° inversion pulse and the subsequent 90° excitation pulse, known as the inversion time (TI). By selecting an appropriate TI, it is possible to null the signal from specific tissues based on their T1 relaxation properties, thereby enhancing contrast or suppressing undesired signal components. Common IR-based sequences include Short TI Inversion Recovery (STIR) designed to suppress fat signal by selecting a TI that coincides with the null point of fat, such that lesions (with increased water content) are made to appear brighter.

2.2.4 Spinal cord MRI

While the preceding sections have introduced the principles of MRI physics, scanner components, and various pulse sequences, applying these principles to image the spinal cord presents a unique set of technical and anatomical challenges. Figure 2.7 shows partially labeled T1w and T2w images of the spinal cord, highlighting how the cord-CSF contrast differs according to the chosen sequence. Given the spinal cord's small cross-sectional dimensions, susceptibility to motion and field inhomogeneities, automatic image analysis becomes particularly challenging, requiring tailored approaches.

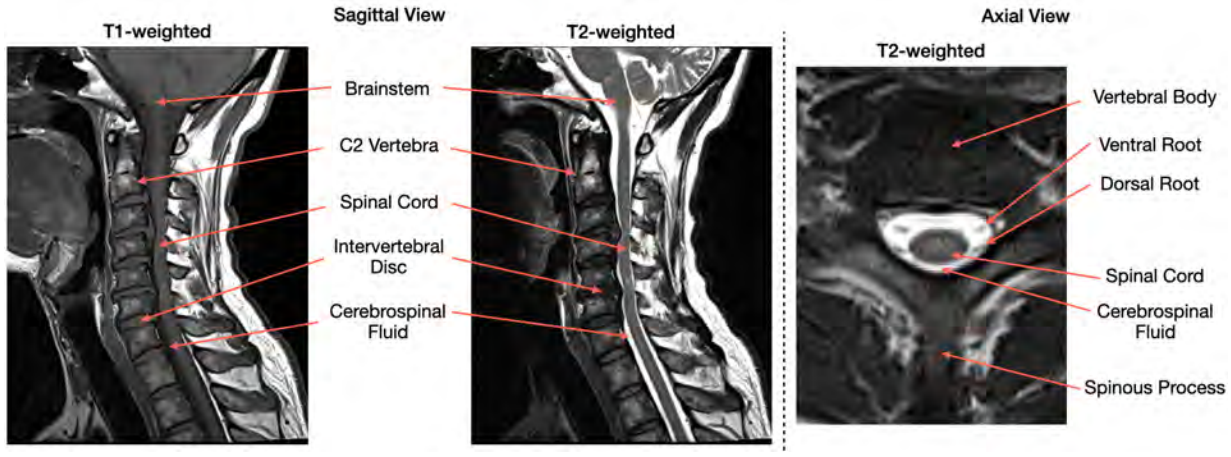

Figure 2.7 Annotated T1-weighted and T2-weighted sequences of the cervical spine (sagittal and axial views). The cord with a cross-sectional diameter of $\sim$ 1 cm is surrounded by the CSF, bones, and air, making it susceptible to field inhomogeneities and motion artifacts.

Partial volume effects

One of the most common phenomenon in neuroimaging is the *partial volume effect* (PVE). It occurs when when a voxel (*i.e.* the smallest unit of a 3D image) contains the signal intensities of multiple tissue types, leading to an inaccurate representation of the tissue proportions within that voxel [48, 51]. This typically happens when the voxel size is comparable to or larger than the size of the structures being imaged. The result is that the voxel intensity is a weighted average of the different tissue types within it, misrepresenting the true tissue composition. Mathematically, in the context of the spinal cord, suppose that a voxel in a cross-sectional slice contains fractional amounts of the cord (w_{SC}) and the CSF (w_{CSF}). Let I_{SC} and I_{CSF} be the true signal intensities of the cord and CSF, respectively. Then, as a consequence of PVE, the MR signal from the voxel I_V is given by:

$$I_V = w_{SC} I_{SC} + w_{CSF} I_{CSF}$$

Figure 2.8 illustrates this phenomenon, where we can notice that the intensity of the voxel at the cord-CSF boundary is averaged with equally-weighted contribution from both tissues. Note here that PVE could worsen with large voxels as one voxel now contains tiny proportions of more tissues.

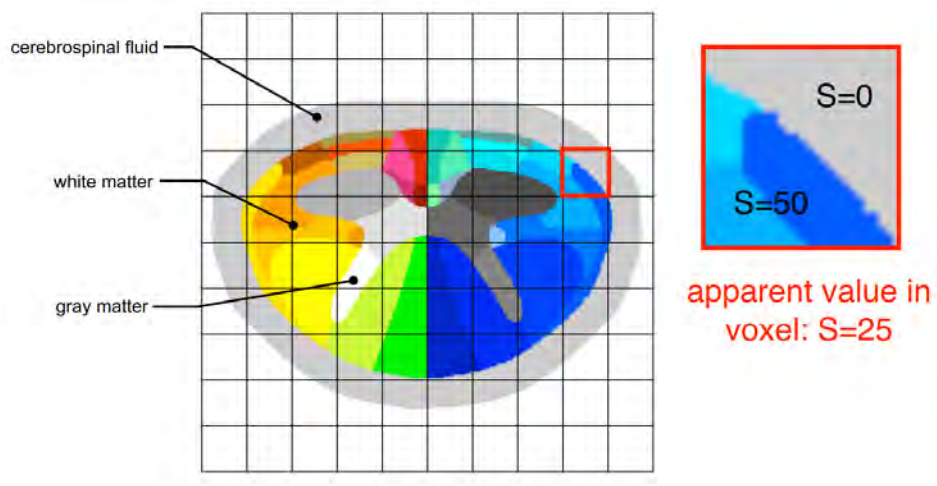

Figure 2.8 Schematic highlighting various tissues in the spinal cord and how partial volume affects their signal intensities. Source: Spinal Cord Toolbox [8].

While PVE mainly occurs due to the limitations of spatial resolution across various MR sequences, high resolution images with small voxel sizes have a higher likelihood of containing a single tissue type, thereby reducing PVE. However, high-resolution acquisitions come at

the cost of low SNR and longer acquisition times as smaller voxels require more time/measurements to accumulate enough signal. Figure 2.10 shows a few examples of sagittal and axial acquisitions with different resolutions. Depending on the MR sequence, the effect of partial volume can be worsened with heavily blurred cord-CSF boundaries (*e.g.* EPI and DWI sequences).

2.2.5 Morphometric measures of the spinal cord

From a neuroimaging standpoint, *morphometry* refers to the quantitative measurement and analysis of the shape, size, and structural properties of the brain/spinal cord and their specific regions, typically derived from MRI images. Morphometric analyses are extremely useful in studying anatomical differences across populations (*e.g.* normative measures in healthy controls and diseased populations), longitudinal monitoring of disease progression and the effects of treatment interventions. Such measures provide insight into micro- and macro-structural changes in the brain and spinal cord improving our understanding of the evolution of certain chronic neurological conditions.

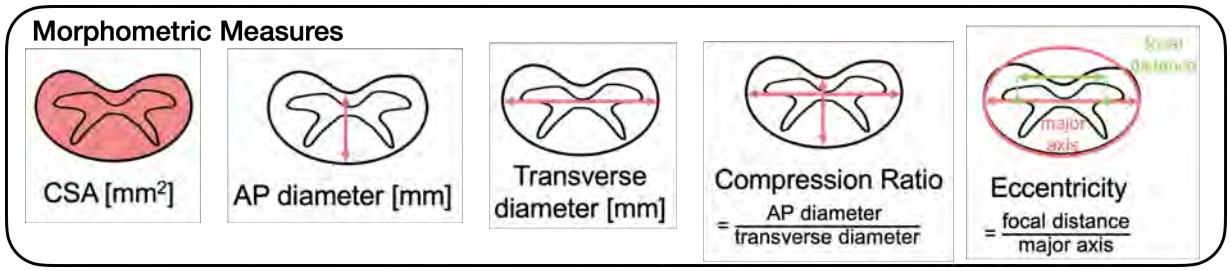

Figure 2.9 Common measures of spinal cord morphometry. Adapted from Valošek *et al.* [9].

Figure 2.9 illustrates some of the common morphometric measures of the spinal cord. CSA quantifies the atrophy of the spinal cord and is computed as the area of the spinal cord in the transverse plane. Precise estimation of CSA is limited by the axial resolution of the image, therefore, CSA is typically measured by averaging over multiple slices across a given set of vertebral levels. The AP diameter measures the diameter of the spinal cord in the anterior–posterior direction, while the transverse diameter measures the cord from side to side (right-left). The compression ratio reflects the flattening of the spinal cord and is defined as the ratio of the AP diameter and the transverse diameter. Treating the cord as an ellipse, eccentricity is computed as the ratio of the distance between the two focal points of the ellipse and the length of the longest diameter (*i.e.* the major axis). The eccentricity of an elongated cord is close to 1, while that of a circular cord is close to 0, indicating the degree of roundness or convexity of the spinal cord.

Quantitative measures of spinal cord morphometry such as the cross-sectional area (CSA) and cord diameter are useful MRI-derived biomarkers used for diagnosing neurological diseases such as MS [102] and DCM [106].

2.2.6 Spinal cord pathologies

After having identified the key structures in spinal cord MRI images and how morphometric measures could serve as important biomarkers for various neurological diseases, we now dive deeper into some of the most common pathologies affecting the cord and the role of MRI in diagnosing them. This section briefly introduces the most prevalent spinal cord pathologies and how image segmentation plays a role in deriving image-based clinical biomarkers from various MRI sequences.

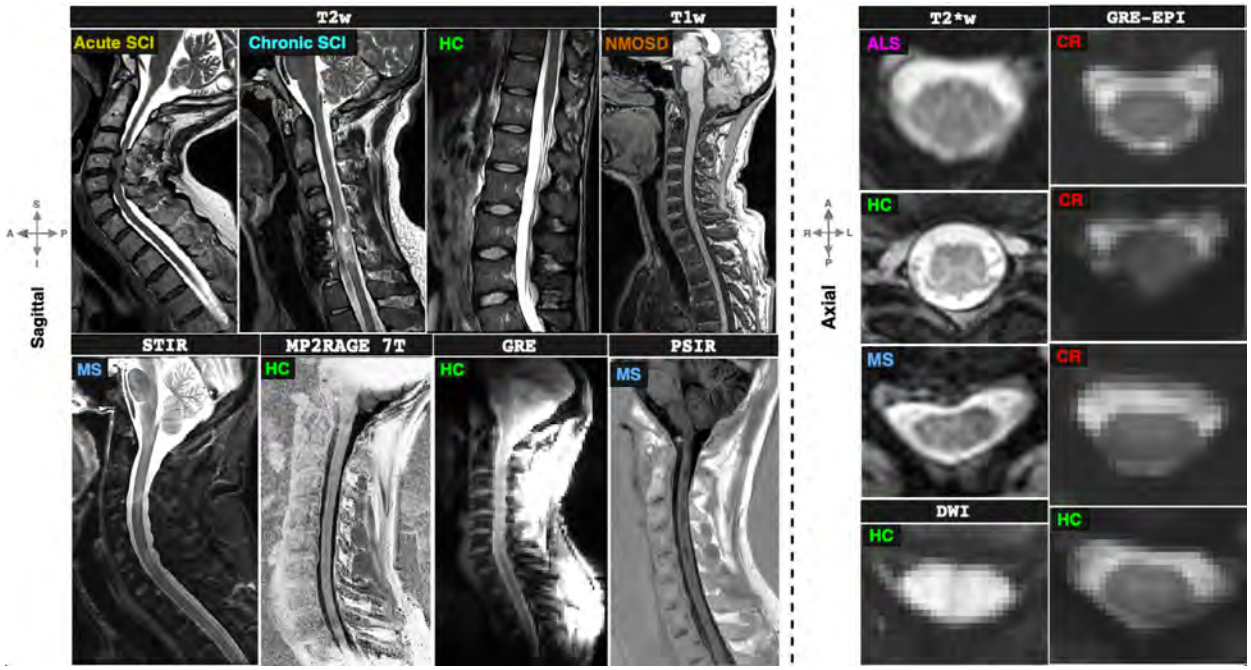

Figure 2.10 Heterogeneity in spinal cord images across various contrasts, pathologies and image resolutions. Legend: SCI: spinal cord injury, DCM: degenerative cervical myelopathy, MS: multiple sclerosis, NMOSD: neuromyelitis optica spectrum disorder, ALS: amyotrophic lateral sclerosis, CR: cervical radiculopathy, and HC: healthy control.

Figure 2.10 gives an overview of the type of images obtained from various pulse sequences, showing representative samples of various spinal cord pathologies. The ability to distinguish the cord (especially its boundaries) from the surrounding CSF depends on the image resolution, with some sequence protocols designed to obtain high resolution images. In addition to the population-level variations in spinal cord length and curvature resulting in variably-sized

images, some pathologies significantly alter the structure of the cord (*e.g.* due to a traumatic injury or age-related compression), contributing to the heterogeneity in spinal cord images.

Spinal cord injury

SCI is defined as damage to the spinal cord that causes temporary or permanent changes in its function, often associated with high-socioeconomic burden (Figure 2.11, left). SCIs are categorized by their origin into two types: *traumatic SCI* stems from an external physical impact (e.g., car crashes, falls, sports injuries, or violence), whereas *non-traumatic SCI* arises from an underlying acute or chronic disease process, including conditions like tumors, infections, or degenerative disc disease. Currently, there is no cure for SCI but rehabilitation has been shown to improve outcome [114]. Not only do SCIs cause damage to the injury site (*i.e.* the primary injury), they also trigger a cascade of complex secondary pathological processes above and below the injury site [115,116], also affecting the brain [117,118]. The primary and secondary injuries in traumatic SCI can be temporally divided into multiple phases: acute ($<$ 48 hours), subacute (48 hours to 14 days), intermediate (14 days to 6 months) and chronic ($>$ 6 months).

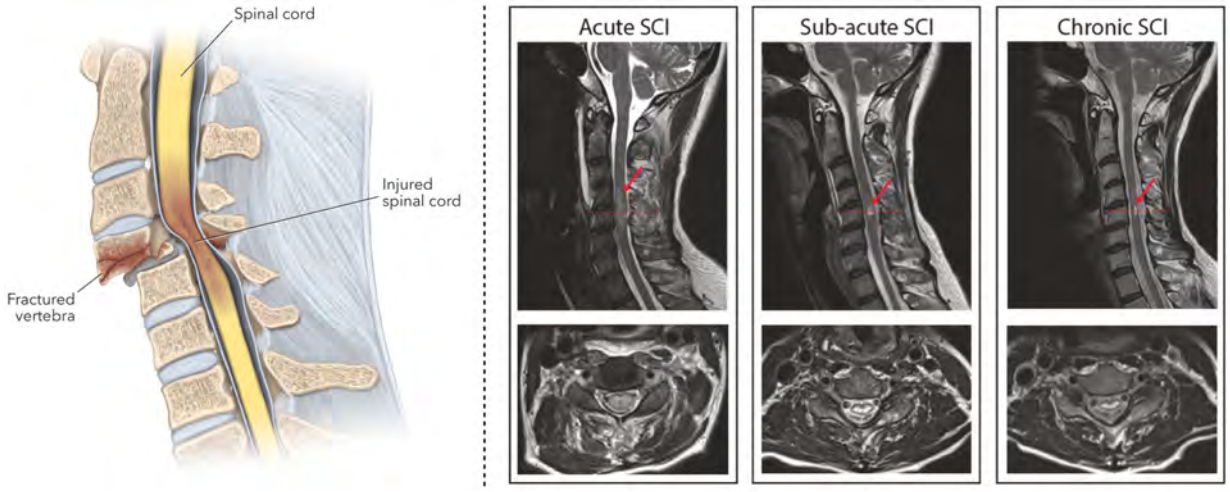

Figure 2.11 Left: Schematic of SCI. Right: Lesion evolution with persisting midsagittal tissue bridges over time in a 63-year old patient with traumatic SCI. Sagittal and axial T2-weighted scans showing lesion evolution in acute (1 day post-SCI), subacute (1 month post-SCI), and chronic phase (24 months post-SCI). Sources: Cohen-Gadol [10] and Seif *et al.* [11].

Pathophysiologically, in the acute phase, the initial mechanical impact triggers a secondary injury cascade characterized by swelling, bleeding, reduced blood flow, inflammatory cell infiltration, and the release of toxic substances, ultimately leading to neuronal and glial cell death, demyelination, and neural circuit disruption. During the subacute phase, ongoing

edema exacerbates ischemia, while persistent inflammation causes further cell death and the formation of microcavities as tissue is damaged. Finally, in the intermediate and chronic phases, axonal degeneration continues, and the astroglial scar matures, becoming a significant barrier to regeneration. Concurrently, cystic cavities merge, further impeding axonal regrowth and cell migration.

Conventional MRI sequences (T1w and T2w) are useful for assessing the level of damage, the extent of intramedullary/extramedullary abnormalities (edema and hemorrhage), degree of spinal cord compression, disc herniation, spared ligaments and paraspinal tissues [34, 119] (Figure 2.11, right). Combined with clinical assessments, MRI-derived biomarkers hold potential in predicting clinical outcome for patients with SCI. For instance, sagittal and axial T2w MRI at the injury level can be used to measure the intramedullary lesion length, lesion volume, tissue bridges, and the extent of spinal cord compression, all prognostic markers for predicting recovery [34, 101].

One of the key challenges in limiting the diagnostic utility of MRI is the presence of image artifacts reducing the reliability of MRI-derived metrics. As most patients with SCI undergo surgery to stabilize the spinal cord, the presence of metallic implants causes substantial image artifacts such as geometric distortion and signal loss, worsening with stronger magnetic fields. Due to the lack of fully automatic methods for deriving MRI-based imaging biomarkers, correlation of the structural changes with clinical examinations is currently done manually.

Degenerative cervical myelopathy

Degenerative cervical myelopathy (DCM), the most common form of non-traumatic SCI, is a condition where age-related changes in the cervical spine (neck) lead to compression of the spinal cord, causing neurological symptoms such as pain and numbness in limbs, poor coordination and imbalance (Figure 2.12, left). DCM can lead to progressive disability and paralysis due to chronic spinal cord compression and non-traumatic SCI [120]. Although traumatic SCI and DCM have different aetiologies, they show similar degrees of spinal cord pathphysiology remote from the injury site, suggesting the involvement of similar secondary degenerative mechanisms. Specifically, the sustained mechanical stress, combined with dynamic factors like neck movement, leads to both direct tissue damage and compromised blood supply (ischemia) to the spinal cord. At a cellular level, this results in the degeneration and death of neurons and oligodendrocytes, impairing nerve signal transmission and leading to demyelination. Over time, these processes contribute to a progressive loss of neural function, manifesting as the characteristic symptoms of DCM.

As in the case with traumatic SCI, structural MRI techniques provide information about

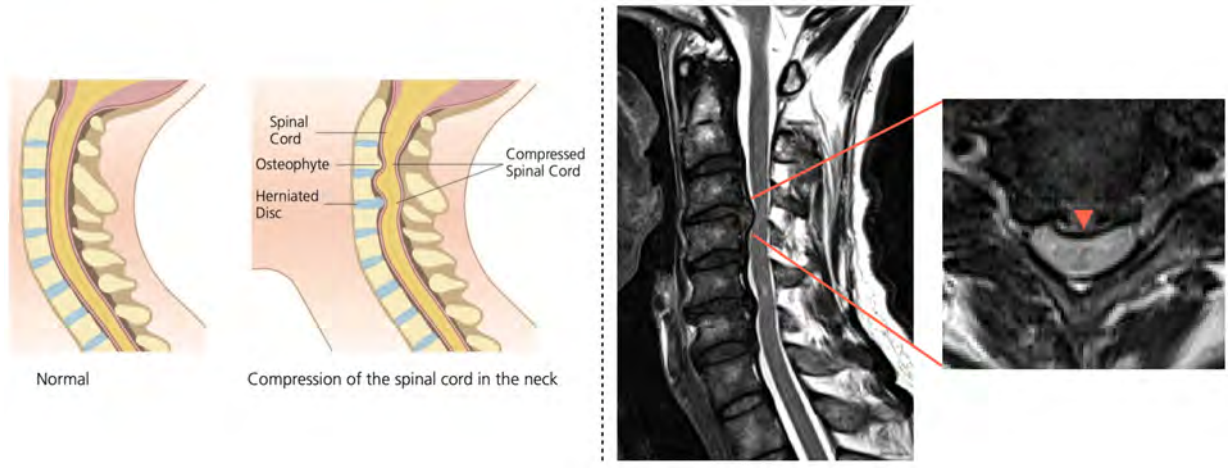

Figure 2.12 Left: Schematic of healthy cervical spine and spine affected with DCM (Source: Mileski [12]). Right: Site of compression on T2-weighted sagittal and axial scans.

the extent/level of injury (compression) and the presence of edema and/or hemorrhage [121] (Figure 2.12, right), aiding in the diagnosis and treatment planning of patients with DCM. Spinal cord segmentation is used to compute morphometric measures (specifically, the cross-sectional area) and longitudinally evaluate the severity of compression by measuring the spinal cord atrophy [106, 122, 123]. However, automatic segmentation is challenging due to the narrowing of the spinal canal from the herniated disc, where existing segmentation methods trained on non-SCI data, fail to delineate compressed cord with distorted shapes.

Multiple sclerosis

Multiple Sclerosis (MS) is a chronic, autoimmune disease of the central nervous system. It is characterized by the immune system attacking the myelin sheath, which protects nerve fibers in the brain and spinal cord, causing inflammation and damage (Figure 2.13, left). This damage generally leads to a variety of symptoms, including vision problems, weakness, numbness, fatigue, and cognitive difficulties [124]. MS can have different courses (categorized into various phenotypes), with some patients experiencing relapses and remissions, while others experiencing a more gradual and progressive neurological decline over time.

MRI is a vital tool for diagnosing and monitoring MS. Like in DCM, MRI scans help in identifying and longitudinally monitoring the areas of demyelination (lesions or plaques) and cord atrophy in the brain and spinal cord. Spinal cord atrophy, particularly in the cervical region, is a recognized imaging biomarker for diagnosis and prognosis of MS, with distinct phenotypes associated with different atrophy rates [102, 104, 105]. It measures tissue loss and

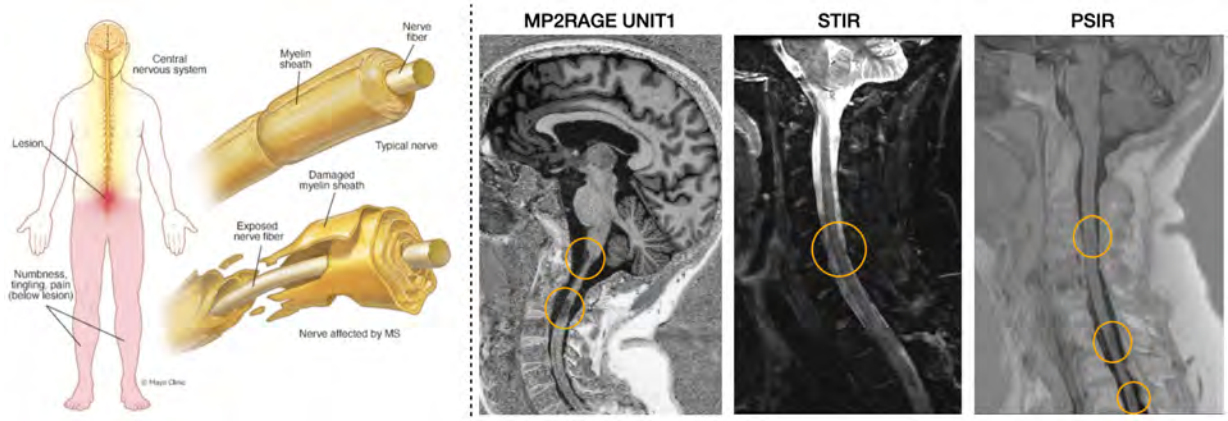

Figure 2.13 Left: Schematic of healthy nerve and nerve with MS along with the common symptoms (Source: Mayo Clinic [13]). Right: MS lesions appearing as hyper-/hypo-intensities highlighted on three different MRI sequences.

can be a valuable indicator of disease progression, disability and response to treatment in various forms of MS. Atrophy is typically measured by segmenting the spinal cord and computing the CSA, with its precision depending on the image resolution and, more importantly, the robustness of the segmentation algorithms. The accuracy of segmentation, especially at the cord boundaries, and whether the segmentation mask is binary (discrete; 0/1 values) or soft (continuous; values between 0-1) have a significant impact on the estimated CSA. For instance, CSA measured across different contrasts should be similar, yet, we will see later in Chapter 5 how this seemingly logical fact does not hold in state-of-the-art segmentation algorithms and how we address this challenge.

2.3 Image Segmentation

Given that MRI provides superior soft tissue contrast and anatomical detail, the quantitative analysis of spinal cord morphology often requires the precise delineation of specific structures from raw image data. Segmentation of these structures is used for deriving morphometric measures such as the CSA and lesion load, thereby bridging the gap between qualitative visual assessment and objective measurement essential for diagnosis and prognosis for pathologies affecting the spinal cord.

The task of semantic segmentation using DL can be defined mathematically as follows [125]: Let $\mathbf{X}$ denote the input space of images. Each image $\mathbf{x} \in \mathbf{X}$ is a discrete representation composed of N pixels, with spatial dimensions (*e.g.*, $W \times H \times D$ for 3D images, where $N = W \times H \times D$). We can represent an image $\mathbf{x}$ by its pixel values, $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$.

The label space is defined as $\mathbf{Y} = \{y_1, y_2, \ldots, y_K\}$, where K is the total number of distinct semantic classes. Pixels that do not correspond to any specific semantic object are assigned to the designated *background* class, $\mathbf{b} \in \mathbf{Y}$. Given an input image $\mathbf{x}$, the objective of semantic segmentation is to assign each pixel $\mathbf{x}_i$ a corresponding label $\mathbf{y}_i \in \mathbf{Y}$, thereby creating a pixel-wise semantic map. The ground truth output space for such maps is therefore defined as $\mathbf{Y}^N$, representing an N-tuple of labels from $\mathbf{Y}$.

Given a training set $\mathbf{T} \subset \mathbf{X} \times \mathbf{Y}^N$, a neural network f_θ with trainable parameters θ is employed to learn the mapping from the input space $\mathbf{X}$. Specifically, f_θ maps an input image $\mathbf{x}$ to a tensor of pixel-wise probability distributions over the semantic classes. Thus, the function mapping can be precisely stated as $f_\theta : \mathbf{X} \to \mathbb{R}^{N \times K}$, where the output for each pixel i (i.e., $(f_\theta(\mathbf{x}))_i \in \mathbb{R}^K$) represents a vector of unnormalized scores (logits) for each of the K classes, which are subsequently transformed into a probability vector (e.g., via a softmax function). The output segmentation mask is obtained by:

$$\mathbf{y}^* = \{\text{argmax}_{c \in \mathbf{Y}} f_\theta(\mathbf{x})\,[i, c]\}_{i=1}^{N},$$

where, $f_\theta(\mathbf{x})\,[i, c]$ denotes the probability of the label of class c for pixel $\mathbf{x}_i$.

2.4 Elements of an Automatic Segmentation Pipeline

2.4.1 Preprocessing

MRI scans are acquired using diverse acquisition protocols and can be noisy and contain artifacts. Data preprocessing steps are always applied to transform raw data into inputs suitable for training segmentation models. Commonly applied techniques include intensity normalization (*e.g.*, standardizing voxel values to a common range to ensure consistent input across different scans) and resampling (adjusting voxel spacing to a uniform resolution to have fixed-size inputs).

Data augmentation Medical datasets are relatively small compared to natural-image datasets and therefore, data augmentation plays a critical role in medical image segmentation as models rely on extensive data augmentation for better generalization capabilities [126]. A wide variety of transformations are applied to the images and their corresponding GT masks, such as geometric transformations (*e.g.*, rotations, flips, scaling, elastic deformations) and intensity transformations (*e.g.*, brightness adjustments, contrast changes, adding Gaussian noise, blurring). These techniques aim to simulate real-world variations in image acquisition and patient anatomy, thereby exposing the model to a wider range of scenarios and improv-

ing its robustness. The impact of data augmentation has been well-studied and interestingly, extensive data augmentation, to the extent of obtaining *unrealistic* images has been shown to improve segmentation robustness and accuracy [69]. Most of the augmentation techniques are implemented in open-source packages such as MONAI and TorchIO.

2.4.2 Model architectures

Several DL architectures have emerged in the last decade [127]. However, there are a handful of architectures that are the cornerstones of DL and help drive the field forward. Among these, Convolutional Neural Networks (CNNs) have been particularly transformative, especially for computer vision tasks like image segmentation and classification. A notable mention for medical image segmentation is the U-Net architecture [128], which, with its encoder-decoder pathway and skip connections, effectively captures both contextual and localized information, leading to highly robust segmentations. Due to their success in medical imaging (and also recently with stable diffusion), several variants of U-Net, specifically emulating the encoder-decoder have emerged.

As new architectures continued to be developed, recent works have shown that novel architectures based on transformers and/or state space models achieve subpar performance compared to well-tuned CNN-based networks when subjected to rigorous validation [98, 129–131]. One can notice a recency bias towards novel architectures despite only obtaining incremental improvements on carefully-curated benchmarking datasets without sufficient evidence on real-world clinical data. Furthermore, novel architectures relying on latest advancements in GPU technology cannot be deployed in clinical settings with modest computational resources. We refer the reader to [127] for a detailed review of model architectures and focus on the U-Net architecture below.

The U-Net network is composed of two parts (Figure 2.14). The first part is the contracting path that employs the downsampling module consisting of several convolutional blocks to extract semantic and contextual features. And in the second part, the expansive path applies a set of convolutional blocks equipped with the upsampling operation to gradually increase the spatial resolutions of the feature maps, usually by a factor of two, while reducing the feature dimensions to produce the pixel-wise classification score. The skip connections copy the outputs of each stage within the contracting path to the corresponding stages in the expansive path. This design choice propagates essential high-resolution contextual information along the network, which encourages the network to re-use the low-level representation along with the high-context representation for accurate localization. The U-Net has become the *de facto* backbone in medical image segmentation since 2015, and several variants of the model have

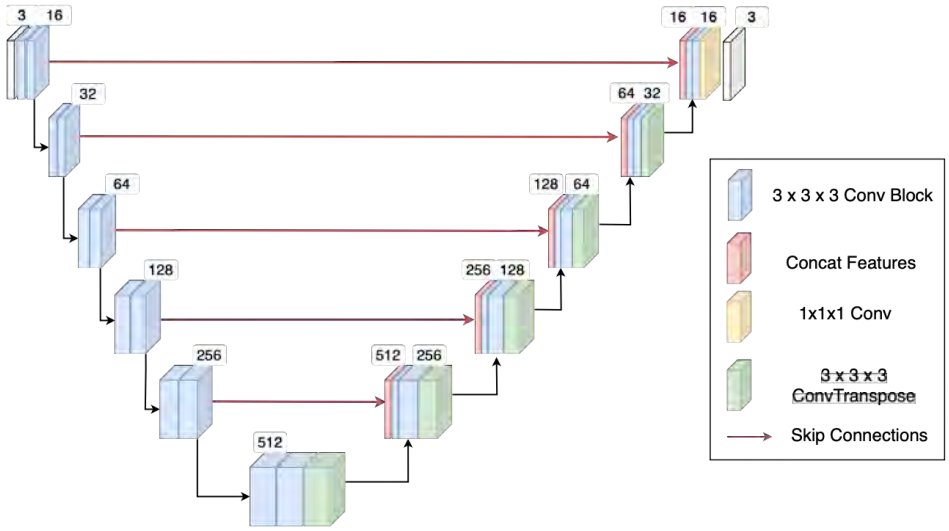

Figure 2.14 3D UNet architecture.

been derived to advance the state-of-the-art based on it [25].

2.4.3 Loss functions

In supervised learning, loss functions quantify a measure of (dis)similarity between ground-truth (GT) segmentations and the model predictions. They play a crucial role in guiding the learning process as model parameters are updated based on the gradients derived from the loss function. As medical image segmentation tasks suffer from the class imbalance problem, where number of voxels in the target object to be segmented are heavily outnumbered by the voxels in the rest of the image (*e.g.* lesions in the spinal cord), choosing the right loss function is important.

Depending on how the dissimilarity is measured, loss functions can be categorized into four types: (i) pixel-based, (ii) overlap-based, (iii) boundary-based, and (iv) combination/compound losses. Pixel-based loss functions operate at the individual pixel level and aim to ensure accurate classification of each pixel within the predefined semantic classes by measuring the deviation of predicted pixels from the corresponding GT pixels. In contrast, overlap-based loss functions prioritize overall class segmentation by maximizing the alignment of target objects between the predicted mask and the GT mask. Boundary-based loss functions optimize for the precise delineation of the object boundaries by minimizing the distance between the prediction and GT masks. As loss functions from these categories do not tackle the class-imbalance problem in isolation, combining these loss functions presents a best-of-all-worlds scenario in mitigating class imbalance. As the most commonly used loss functions belong

to the pixel-based, overlap-based and their combinations, we consider them in greater detail below. Table 2.1 shows the notational convention used for the mathematical definitions of loss functions discussed in this section.

Table 2.1 Notation for the mathematical definition of loss functions

Symbol	Description
N	Number of pixels in the image
C	Total number of target classes to segment {1, ..., C}
y_i^c	Indicator function: 1 if i^{th} pixel belongs to class c; 0 otherwise
y_i	One-hot encoding vector representing the target class of the i^{th} pixel
p_i^c	Predicted probability of i^{th} pixel belonging to class c
p_i	Predicted class probabilities for i^{th} pixel
$p_i \cdot y_i$	Predicted probability of the target class at i^{th} pixel
w	Weight assigned to target classes

Pixel-based Cross-entropy (CE) is one of the most commonly used loss functions. In essence, it is derived from the Kullback-Leibler (KL) divergence between two probability distributions. In segmentation, these refer of the GT distribution (*i.e.* the space where the GT masks lie) and predicted distribution (*i.e.* the space where the model predictions lie), measuring how well the predictions match the GT labels. Once the model outputs the pixel-wise probability maps representing the likelihood of each pixel belonging to a certain class, CE loss is computed using the negative logarithm of the predicted probability for the target class at each pixel. Due to the nature of the log function, the loss approaches 0 and the predicted probabilities for the target classes approaches 1. Mathematically, the loss is defined as:

$$L_{\text{CE}}(p, y) = -\frac{1}{N}\sum_{i=1}^{N}\log(p_i \cdot y_i) = -\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N} y_i^c \ \log(p_i^c)$$

Note, as y_i is a one-hot encoded vector, the loss is only computed over the predicted probability for the target class. Weighted cross entropy is a common extension of the original CE loss, where different weights are assigned to each target class to mitigate class imbalance. Typically, the weights are inversely proportional to the class frequency, meaning that classes with fewer number of voxels have larger weights. The weighted CE loss is given by:

$$L_{\text{CE}}(p, y) = -\frac{1}{N}\sum_{i=1}^{N}(\omega \cdot y_i)\log(p_i \cdot y_i) = -\frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N} \omega_c \ y_i^c \ \log(p_i^c)$$

Loss functions derived from CE tend to focus on overcoming the class imbalance issue with

the topK [132] and focal loss [133] being the notable ones. For instance, focal loss assigns different weights to easy and hard samples. Hard samples are individual voxels or regions that are misclassified with a high probability (*e.g.*, boundary voxels or small objects), while easy samples are those that are correctly classified with a high probability (*e.g.*, background regions). This helps in balancing the influence of easy- and hard-to-segment voxels on the overall loss.

Overlap-based Overlap-based or region-level loss functions zoom out from pixel-level losses and aim to maximize the accuracy of the segmentation at the object-level, capturing the shape of the object in essence. Dice loss [134] is the most popular loss function in this category.

Dice loss directly optimizes for the Dice similarity coefficient, which measures the overlap between two sets of objects (*i.e.*, the predicted mask and the GT mask). Given two sets P and Y, it is computed as the ratio of twice the size of the intersection of the two sets, over the sum of the number of elements in each set: $\text{DSC} = 2\frac{|P \cap Y|}{|P|+|Y|}$, where P is prediction and Y is the GT mask. The Dice loss is the differentiable version of the Dice coefficient adapted for training segmentation models. It is given by:

$$L_{\text{Dice}} = 1 - \frac{1}{C}\sum_{c=0}^{C-1}\frac{2\sum_{i=1}^{N} p_i^c \, y_i^c}{\sum_{i=1}^{N} p_i^c + y_i^c}$$

The loss is computed separately for each class and averaged across all classes to have a single value summarizing the accuracy of model prediction. Unlike in sets with binary 0/1 values, Dice loss can be computed on output probabilities $[0, 1]$ making it differentiable and suitable for gradient descent optimization. Dice loss has the advantage of directly optimizing for the evaluation metric and handles class imbalance well due to its sensitivity towards misclassified voxels. Several variants of the loss have also been proposed for specific applications. For instance, generalized Dice loss [135] uses class weights proportional to the inverse of label frequencies similar to weighted CE. Logarithmic form of Dice loss [136] was introduced to tackle highly imbalanced class distributions, centerlineDice (or, clDice [137]) was proposed to segment tubular structures such as vessels and neurons by computing the intersection of the segmentation masks and their (morphological) skeleta. Other common examples of non-Dice loss functions include IoU loss and Tversky loss.

Compound losses There is no consensus on which category of loss functions is better, however, there is empirical evidence favoring compound loss functions, the most common

being the Dice-CE loss [138]. It has been shown that Dice and CE losses act complementarily, with the former being sensitive towards extreme class imbalance and the latter pushes the model predictions towards the GT distribution by penalizing for high KL divergence [139]. In its simplest form, the DiceCE loss is given by:

$$L_{\text{DiceCE}} = \alpha L_{\text{Dice}} + \beta L_{\text{CE}}$$

where, α and β are respective weights for the loss functions. Because of the popularity of Dice loss in most medical image segmentation tasks, it is also combined with focal and tversky losses.

Regression losses Another category of losses not commonly used in medical imaging are regression loss functions. When trained with a regression loss function, the model is tasked with predicting continuous numerical values (*e.g.*, patient age, survival score, etc.). These loss functions quantify the error between the model's continuous predictions and the actual GT values, using mean-squared error (L2) or mean-absolute error (L1) as the scoring functions. Wing loss [140] and its variants are one of the commonly used regression losses. Note that instead of treating segmentation as a pixel-wise classification problem, regression losses can be used to *regress* values between 0 and 1, resulting in soft outputs. The distinction between soft and binary masks and the potential importance of soft masks on downstream clinical applications is discussed in detail in Chapter 5.

2.4.4 Evaluation metrics

Evaluation metrics measure how well the models perform outside their training data. Rigorous validation on out-of-distribution test data using an appropriate set of evaluation metrics is critical for widespread adoption and clinical utility. This section formally defines the evaluation metrics used in the subsequent chapters of the thesis and discusses some of their common pitfalls. We also make a case for why purely relying on metrics-based comparison is problematic (as is the case with current the landscape of medical AI) and how morphometrically-grounded evaluation can improve the likelihood of clinical adoption.

Spatial overlap-based The metrics in this category can be derived from various combinations of the four basic cardinalites of the confusion matrix: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Following the same notation

as in Table 2.1, they are mathematically defined as:

$$\text{TP} = \sum_{i=1}^{N} \mathbb{I}(Y_i = 1 \text{ and } P_i = 1) \; ; \qquad \text{TN} = \sum_{i=1}^{N} \mathbb{I}(Y_i = 0 \text{ and } P_i = 0)$$
$$\text{FP} = \sum_{i=1}^{N} \mathbb{I}(Y_i = 0 \text{ and } P_i = 1) \; ; \qquad \text{FN} = \sum_{i=1}^{N} \mathbb{I}(Y_i = 1 \text{ and } P_i = 0)$$

where $\mathbb{I}(\cdot)$ is the indicator function that outputs 1 when the inner condition is true else 0.

The Dice similarity coefficient [95] is *the* most commonly used metric in comparing automatic and manual GT segmentations. Other similar metrics include: (i) True Positive Rate (or, *recall*) measuring the proportion of positive voxels in the GT mask that are correctly identified in the prediction as positive, (ii) Positive Predictive Value (or, *precision*), measuring the proportion of positive voxels in the prediction that are part of the GT.

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN} \; ; \quad \text{Precision} = \frac{TP}{TP + FP} \; ; \quad \text{Dice} = \frac{2TP}{2TP + FP + FN}$$

A major limitation of the metrics defined above is that they treat every voxel independently and compute the cardinalities based on the agreement at the voxel level. When segmenting tiny and sparse objects such as lesions, the metric values are heavily influenced by the large number of background voxels (resulting in a high TN count). Thus, the results from pixel-wise metrics are often inflated and appear "good" due to the vast number of correctly classified background voxels.

To mitigate the bias of pixel-wise metrics, lesion-wise (or, instance-wise) metrics were introduced, where the evaluation is done at the object-level by treating the tiny structures as blobs and calculating the above metrics for each blob. The procedure for obtaining lesion-wise metrics involves:

1. **Connected Component Analysis (CCA)**: Applying CCA to both the GT mask (Y) and the predicted mask (P) to identify individual, disconnected blobs (lesions).
2. **Matching**: Establishing a correspondence between the GT lesion and the predicted lesion by defining a minimum overlap threshold.
3. **Counting**: Compute TP, FN and FPs comparing each GT and predicted lesion to get lesion-wise sensitivity, lesion-wise precision and lesion-wise F_1 scores.

Volume-based These metrics quantify the volumetric accuracy by measuring the difference between the volumes of the GT and predicted segmentations. Indirectly, they aim to assess

how well the prediction estimates the size of the overall anatomical structure. Relative volume error (RVE) measures the normalized absolute difference between the predicted and GT volumes, indicating the fractional error in the estimated volume compared to the actual volume. Suppose that P_{vol} and Y_{vol} are the volumes of the prediction and the GT, RVE is defined as:

$$\text{RVE} = \frac{\mid P_{\text{vol}} - Y_{\text{vol}} \mid}{Y_{\text{vol}}}$$

A positive RVE indicates that model predictions tend to over-segment the structure, while a negative RVE implies under-segmentation, and RVE $= 0$ indicates a perfect match in volumetric accuracy.

Spatial distance-based These metrics quantify the geometric discrepancies between the boundaries or surfaces of the predicted and GT segmentations. Instead of comparing overlapping regions, they directly measure how far apart the corresponding points on the surfaces are. Surface distance (SD) is one of the basic metrics for a distance-based evaluation. Mathematically, let $p \in S_p$ be a point on the surface of the prediction. Its surface distance to S_Y is defined as the minimum Euclidean distance from p to any point in S_Y and vice-versa for any given point $q \in S_Y$:

$$d(p, S_Y) = \min_{q \in S_Y}(||p - q||_2) \; ; \quad d(q, S_P) = \min_{p \in S_P}(||q - p||_2)$$

Surface distance, in its raw form, gives the shortest distance from one specific point on a surface to the other surface.

Why conventional metrics are not enough

It is easy to notice a standard trend/recipe in medical image segmentation studies: researchers propose novel architectures and benchmark their approaches on various combinations of publicly-available datasets using same set of evaluation metrics (*e.g.*, Dice coefficient, surface distance, etc.). Their approaches typically result in incremental improvements over existing methods and in most cases the hyperparameters of previous approaches are not well-tuned for the dataset the approach is evaluated on [141, 142]. More importantly, the choice of metrics does not address the nuance of biomedical need [14]. Consider a simple example in Figure 2.15, illustrating how the Dice coefficient does not represent the true accuracy of segmentation multiple, small structures are involved.

As a consequence of this concerning trend, two major issues have arisen: (i) the number of studies calling for rigorous validation and benchmarking is growing rapidly [14, 98, 130, 141,

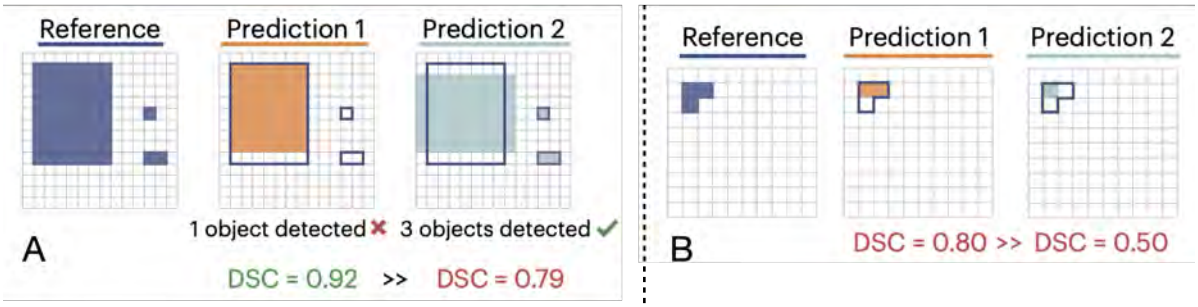

Figure 2.15 Issues with the Dice metric. A) Dice is biased towards single objects and is not suitable for evaluating the segmentation accuracy of multiple objects; B) Dice is susceptible to single-pixel differences can vary significantly, given high inter-rater variability in the annotation of small structures. Adapted from Maier-Hein *et al.* [14].

142] and (ii) despite the wealth of segmentation models claiming to segment a wide-variety of organs across different modalities, they often fail to translate into clinical practice [14, 142].

How can we fix this? While there could be several ways of tackling these issues, one simple way is to take a "bottom-up" approach, where a key clinical biomarker that requires segmentation can be identified and methods to automatically compute the clinical biomarker can then be developed. For instance, in SCI, *tissue bridges*, measuring the width of the spared tissues adjacent to the lesion, are functional biomarkers for recovery in SCI patients. Measuring tissue bridges requires lesion segmentation which are typically obtained by manual annotation. As this process is time-consuming and not scalable to large cohorts, it presents an opportunity to develop SCI lesion segmentation methods (now appropriately grounded with a clinical need) that can robustly solve one particular task and benefit the SCI community. Similar biomarkers exist in MS, where, for instance, CSA of the spinal cord is computed to quantify cord atrophy for longitudinal monitoring of patients. As we shall see in the subsequent chapters, the segmentation tools developed in this thesis are clinically grounded, providing ready-to-use tools for clinicians' needs accessible via the command line interface.

2.5 Binary and Soft Segmentations

Most automatic segmentation algorithms are trained using binary masks with 0/1 values. This black-and-white approach is limiting as it prevents the models from learning subtle intensity differences at the tissue boundaries. In contrast, training with soft masks where the pixels can take on `float` values between $0 - 1$, helps encode levels of uncertainty in models' predictions [143], and potentially account for the aforementioned partial volume

effects occurring at the boundary of tissue interfaces [48, 50, 51]. Figure 2.16 shows a simple illustration of how non-integer values can result in more informative segmentation masks and realistically transition between the cord-CSF interface.

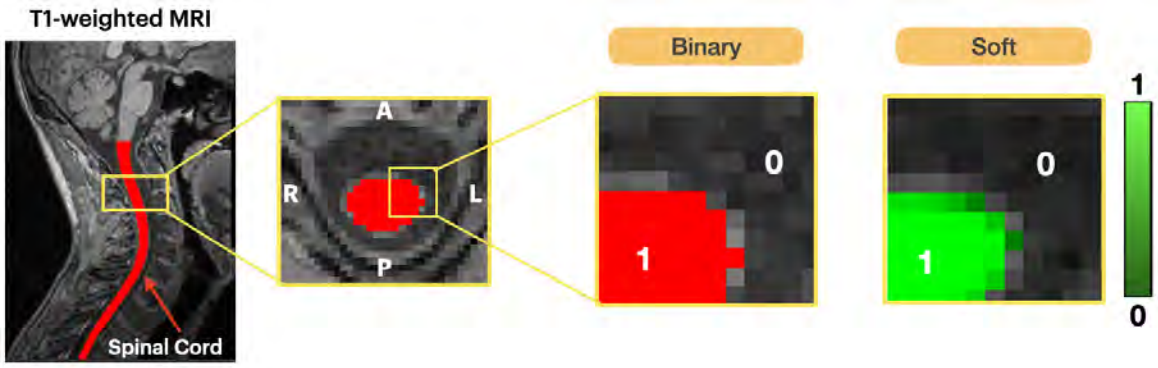

Figure 2.16 Comparison of binary and soft spinal cord segmentations. Notice how soft segmentation shows a gradual transition from the center of the cord to the CSF, whereas, binary segmentation shows an abrupt discontinuation in the cord-CSF interface.

Several studies have reported improved segmentation performance when trained using soft masks and various approaches were used to obtain soft masks. For instance, Kats *et al.* [144] used morphological dilations, where the boundary pixels obtained from the dilated mask were assigned soft values, Li *et al.* [145] obtained soft masks based on signed distances to the annotation boundary and assigned probability values between $0-1$ to them, Gros *et al.* [15] presented a different approach, where soft labels were obtained "for free" by skipping the binarization step after data augmentation. Furthermore, from a clinical standpoint, when soft masks are used for downstream analyses such as mask-based registration or computing morphometrics, the accuracy of a few boundary pixels can significantly impact the outcome.

2.6 Active Learning

DL has become the *de facto* approach for many tasks including segmentation, detection, extraction of clinically-relevant information and computer-aided diagnosis. State-of-the-art models, however, rely on huge corpora of annotated training data and still suffer from generalization issues, especially on rare instances of pathologies occurring on specific anatomies and acquired on contrasts not commonly seen in the public domain. While the costs of labeling medical data are still astronomical, advances in DL techniques such as active learning (AL) and human-in-the-loop computing have helped overcome labeling costs especially with model-assisted labeling [57, 60, 146–148]. Figure 2.17 shows a simple human-aided AL

scenario, where a subset of unlabeled data is annotated and the model is incrementally improved. Note that with each AL phase, as the model gets better in automatic segmentation, the cost of annotation is also reduced.

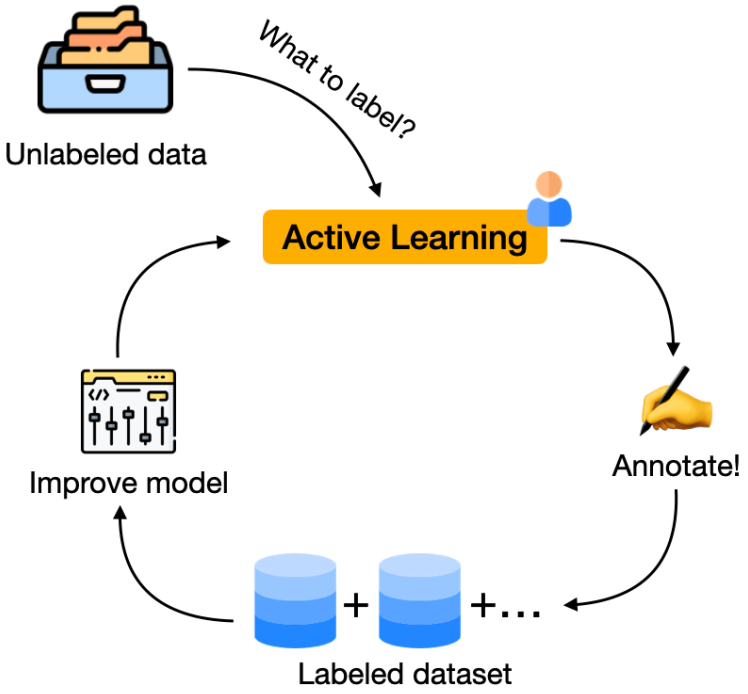

Figure 2.17 Cyclic workflow of human-aided active learning, producing better models more efficiently with a clever selection of samples to label.

Active learning can be formally defined as follows: Let U be a large unlabeled dataset and we have access to an oracle (*i.e.* expert rater) or a group of oracles, who can annotate an unlabeled sample $\mathbf{x}_U$ an add it to the labeled dataset D. The goal is to train a model $f_\theta(\mathbf{x} \mid D^*)$, where $D^* \subseteq D$. The conventional, non-AL approach would be to get annotations for every $\mathbf{x}_U$ such that $D^* = D$. However, such as approach is impractical given high labeling costs. In theory, there exists some D^* such that the performance of the trained model $f_\theta(\mathbf{x} \mid D^*) \simeq f_\theta(\mathbf{x} \mid D)$, meaning that a model trained a curated subset D^* achieves similar performance to that trained on the full dataset D. AL approaches aim to find this optimal subset D^* given the current model $f'(\mathbf{x} \mid D')$, where f' and D' are intermediate annotated dataset and model, respectively, at a given phase and U is the unlabeled dataset.

Typically, samples $\mathbf{x}_U$ are selected based on a notion of *informativeness* of the output $f'(\mathbf{x}_U \mid D')$ and a *query* type [57]. The input query is essentially the strategy used to select unlabeled samples. Examples of query types include, stream-based sampling, where each incoming $\mathbf{x}_U$ is picked for labeling, and pool-based sampling, where a batch of data points are sampled

from the unlabeled dataset U and top-N samples are labeled based on their informativeness. There are various ways of defining the criterion. Most commonly-used measures are:

- **Uncertainty-based sampling**: Most uncertain samples are picked for labeling with the assumption that the *information gain* is highest within these samples,
- **Representativeness**: Labeling only the most uncertain samples might skew the model towards a particular distribution. Sampling based on representativeness encourages the model to pick samples from different parts of data distribution thus increase dataset diversity with each AL phase.
- **Random sampling**: Interestingly, random selection of samples performs competitively or better than sophisticated AL strategies, while being computationally inexpensive and easy to implement.

In the context of this thesis, we focus on the AL scenarios where humans (*i.e.*, expert raters) are involved in the sample selection, labeling and interpretation/quality control (QC). Considering the specific example of segmenting spinal cord lesions given a large unlabeled dataset, a first pass over the dataset with visual QC can identify noisy images, images with artifacts, large/tiny lesions. A random selection of such samples can be used for fully manual annotation (D^*). Then, once an intermediate model (f') is trained, the predictions of the model may not be accurate requiring minor additional corrections by expert raters (making it a semi-automatic approach for data annotation) before adding it to the pool of annotated samples for the next phase of training.

In summary, when segmentation tasks require huge amounts of manual labeling (typically over several slices and hundreds of samples), iterative refinement techniques such as human-in-the-loop AL can help reduce the labeling costs and train large-scale models on diverse multi-site datasets.

2.7 Lifelong Learning

The dynamic, ever-evolving nature of medical data presents a critical challenge for the generalizability of DL methods to new, unseen data distributions. While the lack of accessibility and availability of large training datasets is a challenge, existing datasets are also heterogeneous due to variability in imaging protocols, acquisition parameters, scanner manufacturers, and patient demographics.

Model generalization has been tackled using domain adaptation [149] and transfer learning [61, 150] approaches. Domain adaptation aims to *adapt* a model trained on a source

domain (*e.g.*, a dataset from one hospital) so that it performs well on an unseen target domain (*e.g.*, a dataset from a different hospital) where labeled data might be scarce or absent. It focuses on reducing the discrepancy in data distributions between the source and target domains, enabling the model to effectively generalize across these shifts. Transfer learning, a broader concept, aims to leverage knowledge gained from a source task (*e.g.*, classification on ImageNet) to improve performance on a target tasks/domains (*e.g.*, medical image segmentation) that typically have limited labeled data. It involves taking a pre-trained model and fine-tuning it on the new, related task.

Both approaches operate under specific constraints and come with limitations. Domain adaptation typically assumes that while the marginal distributions of data ($P(\mathbf{X})$) may differ between source and target domains, the conditional distribution of labels given data ($P(\mathbf{Y}|\mathbf{X})$) remains largely consistent, or that a mapping can be learned to align these distributions. A key limitation here is its effectiveness relies on the degree of similarity between the source and target domains; if the domain shift is too large or complex, current approaches may struggle to bridge the gap. Transfer learning, on the other hand, assumes that the low-level features learned by model are generic enough to be transferable to the target task. Furthermore, such approaches are susceptible to negative transfer, meaning that the pre-trained weights hurt the performance if the target and source domains are too dissimilar.

While domain adaptation and transfer learning approaches aim to leverage existing knowledge and bridge domain gaps, they often address *one-off* shifts between distinct source and target datasets. However, real-world medical applications frequently encounter a continuous stream of new data, requiring models to adapt and generalize to new pathologies and distribution shifts. Ideally, one can envision a DL model trained to learn the way humans do, that is, learn *continually* by acquiring, retaining, and transferring knowledge across an endless sequence of tasks, while not completely forgetting previous tasks. Note that under the conventional definition of lifelong learning, it is assumed that the learner does not have access (or, has partially-restrained access) to past data [88]. Under such circumstances, maintaining consistent performance on previously-learned tasks when the knowledge is being accumulated to solve new tasks, becomes extremely challenging, leading to a well-known phenomenon called *catastrophic forgetting* [89].

A typical lifelong learning (LL) setup can be defined as follows [88]: Let there exist a sequence of tasks, where each task t represents a set of unique classes $C^{(t)}$, where $C^{(t)} \subseteq \mathbf{Y}$ (the set of all possible classes). The tasks arrive one-at-a-time in a sequence and each task t comes with its set of training data $D^{(t)} = \{(\mathbf{x}_i, \mathbf{y}_i)_{i=s}^{s+n_t}\}$ with n_t samples, where $\mathbf{x}_i \in \mathbf{X}$ and $\mathbf{y}_i \subseteq C^{(t)}$. The output space $\mathbf{Y}^{\mathcal{T}}$ keeps expanding whenever a new task $\mathcal{T}$ is introduced $\mathbf{Y}^{\mathcal{T}} = \cup_{t=1}^{\mathcal{T}} C^{(t)}$.

The learner's goal is still to learn the function mapping the input space to output space across all seen tasks $f_{\mathcal{T}} : \mathbf{X} \rightarrow \mathbf{Y}^{(t)}$.

In addition to the conventional definition of LL, there could also be a slightly different interpretation; that of a *lifelong learning in production* setup, where, a deployed model is continuously updated to maintain its performance over time. This involves periodically retraining the model as new data become available, often to adapt to distributional shifts occurring with real-world clinical data. Crucially, in this scenario, full access to past data might be available and can be leveraged. This often means the model may be trained from scratch on an accumulating dataset (combining old and new data), or just fine-tuned on the growing dataset. While both interpretations aim for models that can adapt and evolve, the conventional LL focuses on overcoming storage constraints and catastrophic forgetting in a sequential, resource-limited settings, whereas the LL models in deployment emphasize maintaining robustness over time without data/storage constraints.

2.8 Summary

In this chapter, we have introduced several topics and concepts presented later in this thesis. We started with the spinal cord, understanding its anatomy and structure and briefly looked at spinal tracts carrying motor and sensory information. We then looked at how the spinal cord is imaged using MRI, understanding the principles of MRI, various components of an MRI scanner, and the resulting pulse sequences. We covered commonly used MRI sequences (*e.g.*, spin-echo, gradient-echo, echo planar imaging, etc.) obtained from the specific temporal application of RF pulses and field gradients. We identified key structures of interest in a cervical scan of the cord and discussed the partial volume effect phenomenon. We then briefly looked at a few common morphometric measures of the spinal cord, how segmentation is a prerequisite to obtain quantitative measurements, and an overview of main pathologies including SCI, DCM and MS. Moving on to the technical part, we mathematically defined the task of semantic segmentation and understood the elements of an automatic segmentation pipeline (*i.e.*, preprocessing, architectures, loss functions and evaluation metrics). Lastly, we introduced the concepts of active learning and lifelong learning, grounding it to how they have been applied in the contributions of this thesis.

CHAPTER 3 RESEARCH OBJECTIVES

In light of the challenges in spinal cord image analysis and issues with evaluation strategies in current medical image segmentation studies, the overarching theme of this thesis is to **develop generalizable, automatic tools for segmenting the spinal cord and lesions across MRI contrasts and pathologies for better estimation of imaging biomarkers**. These issues are tackled in three research objectives: starting with a contrast-specific method proposed to segment lesions in spinal cord injury, moving towards a generalist, contrast-agnostic approach for spinal cord segmentation, and concluding with a framework for continuous training for spinal cord segmentation in a lifelong-learning-in-production scenario. All the tools and methods described in this thesis are open-source and freely accessible from the command-line interface using the Spinal Cord Toolbox [86] package. The research objectives are detailed as follows:

1. Objective 1: **Develop an open-source, automatic tool for the segmentation of T2w intramedullary lesions in spinal cord injury**. Despite the advances in the image analysis of the spinal cord, robust methods for extracting quantitative MRI-derived biomarkers are still lacking. As a result, biomarkers shown to have positive correlations in the diagnosis and recovery in SCI patients are computed manually. Therefore, in this objective, we introduce **SCIseg**, a tool for lesion segmentation in spinal cord injury, aiming to automatically measure clinically-relevant biomarkers such as the lesion volume, intramedullary lesion length, and tissue bridges. As human-annotated datasets in SCI are difficult to obtain, we gathered heterogeneous data from three clinical sites and explored human-in-the-loop active learning strategy to incrementally annotate the datasets for training. Our results showed that the MRI biomarkers measured from SCIseg predictions and manually-annotated masks had no significant difference, proving the reliability of automatic predictions. Since the release of SCIseg, subsequent studies have relied on automatic segmentations for measuring tissue bridges on unseen, external datasets [151] and studying the relation between spinal tract damage and the development of spastic muscle tone in SCI patients [152], demonstrating the clinical utility of SCIseg. The article presented in Chapter 4 addresses our first objective.

2. Objective 2: **Develop an automatic tool for contrast-agnostic soft segmentation of the spinal cord**. Given the rising importance of the spinal cord in understanding various neurological disease mechanisms, obtaining robust spinal cord segmentation is crucial for several downstream clinical tasks. While the previous objective focused on

a single contrast, we focused on scaling to multiple contrasts in this objective and developed a contrast-agnostic spinal cord segmentation model. As existing contrast-specific models lead to biased estimation of spinal cord morphometrics, specifically due to variability in segmentations across contrasts, we designed a novel preprocessing framework to generate unique, soft masks for each contrast and train a segmentation model directly on these soft masks. Comparing with various domain generalization methods and contrast-specific models, we demonstrated that impact of training on soft masks resulting in reduced morphometric variability across contrasts. The article presented in Chapter 5 achieves our second objective.

3. Objective 3: **Design a lifelong learning framework to continually enrich the spinal cord segmentation models on new contrasts and pathologies while automatically monitoring for performance drifts across various model versions**. Medical datasets are hardly ever static. Given the dynamic and chronic nature of certain neurological conditions, data are continually acquired requiring models developed on static, timestamped versions of datasets to continually evolve and generalize to new, unseen data over time. Therefore, in this objective, we designed a framework enabling continuous training of spinal cord segmentation models using human-in-the-loop active learning. As model performance drift is inevitable under constantly-shifting data distributions, we leveraged continuous integration and continuous delivery (CI/CD) software development practices to automatically monitor for drifts in morphometric variability across contrasts. The article presented in Chapter 6 addresses our third objective.

CHAPTER 4 ARTICLE 1: SCISEG: AUTOMATIC SEGMENTATION OF INTRA-MEDULLARY LESIONS IN SPINAL CORD INJURY ON T2-WEIGHTED MRI SCANS

Authors Enamundram Naga Karthik[*,1,2], Jan Valošek[*,1,2,3,4], Andrew C. Smith[5], Dario Pfyffer[6,7], Simon Schading-Sassenhausen[6], Lynn Farner[6], Kenneth A. Weber II[7], Patrick Freund[6,8], Julien Cohen-Adad[1,2,9,10]

Affiliations

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

[*] Shared co-first authorship - authors contributed equally

This article has been published as an open-access article [153] at the *Radiology: Artificial Intelligence* journal on 2024-11-06 as:

> Naga Karthik, E., Valošek, J., Smith, A. C., Pfyffer, D., Schading-Sassenhausen, S., Farner, L., ... & Cohen-Adad, J. (2024). SCIseg: Automatic segmentation of intramedullary lesions in spinal cord injury on T2-weighted MRI scans. Radiology: Artificial Intelligence, 7(1). https://pubs.rsna.org/doi/full/10.1148/ryai.240005

Contributions

Naga Karthik led the project, proposed the active learning approach, trained segmentation models, and wrote the paper. Jan Valošek assisted in the literature review, ran statistical analyses, generated correlation plots between clinical measures and MRI biomarkers, and implemented the automatic measurement of tissue bridges. Together, Naga Karthik and Jan Valošek curated the dataset, performed quality control and manual corrections during the active learning phases, and discussed the implementation of quantitative lesion biomarkers in SCT's `sct_analyze_lesion` function. Andrew Smith, Kenneth Weber and Patrick Freund shared the data. Dario Pfyffer and Simon Schading-Sassenhausen provided the initial versions of the manually-annoated SCI lesion masks. Dario Pfyffer was also involved in the discussion of measurements of tissue bridges. Julien Cohen-Adad provided supervision throughout the project. All co-authors reviewed the draft of the manuscript.

Key Points

1. The SCIseg deep learning model was developed for segmentation of the spinal cord and intramedullary lesions using a multicenter dataset of 191 patients with spinal cord injury who underwent T2-weighted MRI examinations using various scanner manufacturers and acquisition parameters.

2. SCIseg achieved a mean Dice score of 0.92 ± 0.07 (SD) and 0.61 ± 0.27 for spinal cord and spinal cord injury lesion segmentation, respectively.

3. There was no evidence of a difference between quantitative MRI biomarkers, namely, lesion length (P = .42) and maximal axial damage ratio (P = .16), computed from manually annotated lesions and the lesion segmentations obtained using SCIseg.

Summary

The proposed deep learning model accurately segmented the spinal cord and spinal cord injury lesions in a diverse, multicenter dataset of T2-weighted MRI scans.

An extension of this work has been published in the MICCAI 2024 Applications of Medical Artificial Intelligence (AMAI) workshop as:

> Karthik, E. N., Valošek, J., Farner, L., Pfyffer, D., Schading-Sassenhausen, S., Lebret, A., ... & Cohen-Adad, J. (2024, October). SCIsegV2: a universal tool for segmentation of intramedullary lesions in spinal cord injury. In International Workshop on Applications of Medical AI (pp. 198-209). Cham: Springer Nature Switzerland.

Abstract

Purpose: To develop a deep learning tool for the automatic segmentation of the spinal cord and intramedullary lesions in spinal cord injury (SCI) on T2-weighted MRI scans.

Materials and Methods: This retrospective study included MRI data acquired between July 2002 and February 2023. The data consisted of T2-weighted MRI scans acquired using different scanner manufacturers with various image resolutions (isotropic and anisotropic) and orientations (axial and sagittal). Patients had different lesion etiologies (traumatic, ischemic, and hemorrhagic) and lesion locations across the cervical, thoracic, and lumbar spine. A deep learning model, SCIseg (which is open source and accessible through the Spinal Cord Toolbox, version 6.2 and above), was trained in a three-phase process involving active learning for the automatic segmentation of intramedullary SCI lesions and the spinal cord. The segmentations from the proposed model were visually and quantitatively compared with those from three other open-source methods (PropSeg, DeepSeg, and contrast-agnostic, all part of the Spinal Cord Toolbox). The Wilcoxon signed rank test was used to compare quantitative MRI biomarkers of SCI (lesion volume, lesion length, and maximal axial damage ratio) derived from the manual reference standard lesion masks and biomarkers obtained automatically with SCIseg segmentations.

Results: The study included 191 patients with SCI (mean age, 48.1 ± 17.9 [SD] years; 142[74%] male patients). SCIseg achieved a mean Dice score of 0.92 ± 0.07 and 0.61 ± 0.27 for spinal cord and SCI lesion segmentation, respectively. There was no evidence of a difference between lesion length ($P = .42$) and maximal axial damage ratio ($P = .16$) computed from manually annotated lesions and the lesion segmentations obtained using SCIseg.

Conclusion: SCIseg accurately segmented intramedullary lesions on a diverse dataset of T2-weighted MRI scans and automatically extracted clinically relevant lesion characteristics.

Keywords Spinal Cord, Trauma, Segmentation, MR Imaging, Supervised Learning, Convolutional Neural Network (CNN)

4.1 Introduction

Spinal cord injury (SCI) refers to damage to the spinal cord due to traumatic or non-traumatic processes. Traumatic SCI results from acute damage to the spinal cord due to external physical factors [34, 115]. Most patients with traumatic SCI sustain permanent neurological deficits such as motor and autonomic dysfunction, with devastating physical and social consequences [115]. Degenerative cervical myelopathy (DCM), the most common form of non-traumatic SCI, originates from chronic mechanical compression of the spinal cord [121]. While relatively less common than traumatic lesions, ischemic SCI lesions represent up to 20% of all non-traumatic lesions [154, 155] and show a similar course of recovery to traumatic SCI [101, 156].

MRI scans of patients with SCI provide macrostructural information about the level of injury and intramedullary abnormalities (e.g., edema and hemorrhage), and allow the evaluation of soft tissue structures [115, 121] . Importantly, MRI-derived quantitative biomarkers, such as intramedullary lesion length and lesion volume, have demonstrated associations with the neurological prognosis of patients with traumatic SCI [101,103,157–160]. Particularly, smaller lesion length and area were significantly associated with better recovery of patients [101, 158, 159].

Despite recent advances in the automatic processing of spinal cord MRI [86, 120, 161, 162], robust methods for detecting quantitative MRI biomarkers in SCI are lacking. As a result, most studies involve manual identification of these biomarkers [100, 101, 103, 109, 163, 164]. This is a time-consuming task further hampered by inter-rater variability, making it impractical in clinical trials [121]. Furthermore, segmentation of intramedullary SCI lesions on MRI scans poses an extremely challenging task mainly due to the evolving appearance of lesions in different injury phases (e.g., acute, sub-acute, intermediate) [34, 115]. Surgical implants in postoperative MRI scans might also cause severe image artifacts. Deep learning (DL) can improve diagnosis and prognostication in SCI by automating the lesion annotation process, thereby reducing rater-specific biases and facilitating the analysis of large SCI cohorts across sites [165–167]. Indeed, quantitative SCI lesion biomarkers derived from DL-based automatic segmentations have been shown to correlate well with clinical measures of motor impairment [168]. Despite its numerous potential advantages, DL has not been sufficiently explored in the context of SCI [166], with no open-source methods existing to date. This suggests the need for an automatic biomarker identification method that takes into account the complex pathophysiology of patients with SCI, generalizes to multiple sites, and is easily accessible by researchers.

The purpose of this study was to develop an open-source DL-based tool, **SCIseg**, for the automatic segmentation of the spinal cord and intramedullary lesions on T2-weighted MRI scans of patients with SCI. Model-derived quantitative MRI biomarkers of SCI, such as lesion volume, lesion length, and maximal axial damage ratio, were compared with biomarkers derived from manual segmentations. Further, correlation analyses were conducted between biomarkers derived using the automatic SCIseg and clinical scores, specifically pinprick, light touch, and lower extremity motor scores.

4.2 Materials and Methods

4.2.1 Study design and patients

This retrospective study included 191 patients with SCI who underwent MRI at three sites (Balgrist University Hospital Zurich, Zurich, Switzerland; Craig Hospital, Englewood, Colorado, USA; Pitié-Salpêtrière University Hospital; Paris, France) between July 2002 and February 2023. All patients provided written informed consent following Institutional Review Board approval and the Declaration of Helsinki. The inclusion criteria were traumatic, ischemic, or hemorrhagic SCI, presence or absence of surgical hardware, and clinical data available for analyses. Exclusion criteria were concurrent traumatic brain injury beyond concussion and significant pre-existing neurological history (i.e., multiple sclerosis, transverse myelitis, cerebrovascular stroke). Patients from site 2 were clinically assessed using the international standards for the neurological classification of SCI (ISNCSCI) protocol [169] to obtain light touch, pinprick, and lower extremity motor scores, as previously described in studies by Smith et al. [103, 170]. All 191 patients from the three sites were reported previously [100, 101, 103, 170, 171]. These articles used manually annotated lesion masks to study the clinical consequences of SCI and their predictive relationships with motor and sensory functions. In contrast, our study presents a DL-based tool to segment intramedullary SCI lesions automatically.

4.2.2 MRI data and reference standard

The MRI scans were converted from DICOM to NIfTI format and organized according to the Brain Imaging Data Structure (BIDS) standard [172] at individual sites. During this curation process, all sensitive patient information was deleted. T2-weighted (T2w) MRI scans with varying lesion etiologies (traumatic, ischemic, hemorrhagic), injury chronicity (sub-acute, intermediate, and chronic), orientations (sagittal and axial), and voxel sizes were used for this study (Figure 4.1, Table 4.1). Lesions appearing as T2w signal abnormalities

(hyperintense or hypointense voxels corresponding to primary contusions, secondary cytotoxic edema or hemorrhage) were manually annotated as a single object by two raters from site 1 (D.P. and L.F.), one rater from site 2 (A.C.S.), and by two raters (E.N.K. and J.V.) for site 3 using JIM (V.7.0., Xinapse Systems, Aldwincle, UK) and FSLeyes v1.9.0 image viewers. As obtaining the reference standard spinal cord segmentation masks using a fully manual approach is time-consuming, `sct_deepseg_sc` [17] was used to initially segment the spinal cord of patients from all 3 sites, followed by manual corrections wherever necessary. This semi-automatic approach was also used in previous studies [17].

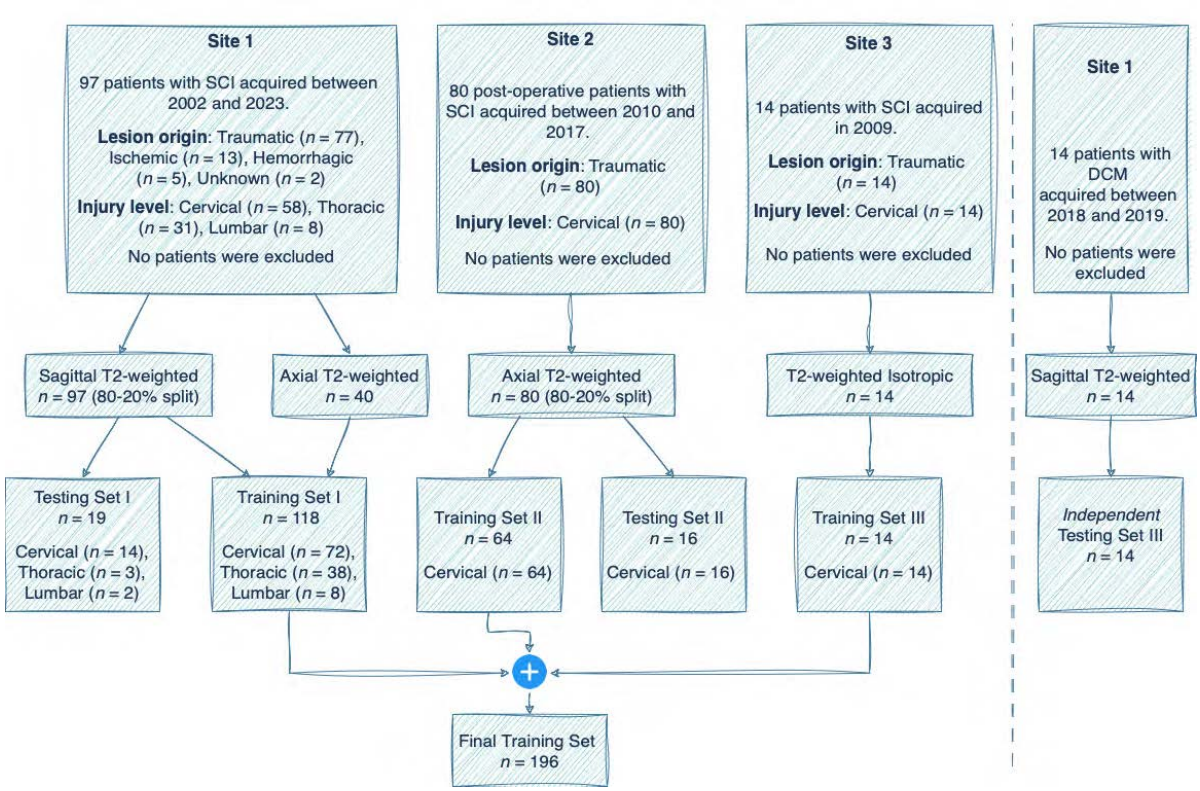

Figure 4.1 **Study Flowchart**. The data included patient cohorts from three sites with heterogeneous image resolutions, orientations, and lesion etiologies. The validation set is included within the final training set. Models were evaluated independently on the test sets of Site 1 and Site 2, along with their evaluation on an external test set of patients with degenerative cervical myelopathy (DCM). Please refer to Table 1 for details on the MRI vendors and field strengths. SCI = spinal cord injury.

4.2.3 Deep learning training protocol

The model was trained in three phases (Figure 4.2). In the initial phase, a baseline segmentation model was trained using a labeled dataset of 78 patients with T2-weighted sagittal

Table 4.1 Characteristics of the study patients

	Site 1	Site 2	Site 3
Number of patients	97	80	14
Number of MRI scans	137[†]	80	14
Sex (male/female)	66/25[*]	65/15	11/2
Age (y) (Mean ± SD)	51.0 ± 19.1	45.8 ± 16.4	42.9 ± 16.7
Age range (y)	17 – 83	15 – 81	21 – 65
Days from injury to MRI (Mean ± SD)	376.2 ± 1364.4[‡]	84.0 ± 212.1	579.1 ± 714.1
Days from injury to MRI (Median)	41.5	21.0	407
Number of patients with surgical implants/hardware	61 3T ($n = 19$), 1.5T ($n = 41$), 1T ($n = 1$)	80 3T ($n = 21$), 1.5T ($n = 59$)	8 3T ($n = 10$)
Lesion origin	Traumatic ($n = 77$) Ischemic ($n = 13$) Hemorrhagic ($n = 5$) Unknown ($n = 3$)	Traumatic ($n = 80$)	Traumatic ($n = 14$)
Injury level	Cervical ($n = 58$) Thoracic ($n = 31$) Lumbar ($n = 8$)	Cervical ($n = 80$)	Cervical ($n = 14$)
Number of patients in train set	Cervical ($n = 44$) Thoracic ($n = 28$) Lumbar ($n = 6$) `Total` ($n = 78$)	Cervical ($n = 64$)	Cervical ($n = 14$)
Number of patients in test set	Cervical ($n = 14$) Thoracic ($n = 3$) Lumbar ($n = 2$) `Total` ($n = 19$)	Cervical ($n = 16$)	0
MRI manufacturers	Siemens ($n = 91$), GE ($n = 5$), Philips ($n = 1$)	Siemens ($n = 20$), GE ($n = 60$)	Siemens ($n = 14$)
MRI field strength	3T ($n = 37$), 1.5T ($n = 59$), 1T ($n = 1$)	3T ($n = 21$), 1.5T ($n = 59$)	3T ($n = 14$)
MRI Sequence Parameters	SAGITTAL T2-weighted: voxel size 0.34×0.34 mm^2 to 0.96×0.96 mm^2; slice thickness 2.2 mm to 4.8 mm AXIAL T2-weighted: voxel size 0.35×0.35 mm^2 to 0.78×0.78 mm^2; slice thickness 1.0 mm to 7.0 mm	AXIAL T2-weighted: voxel size 0.31×0.31 mm^2 to 0.78×0.78 mm^2; slice thickness 3.0 mm to 6.0 mm	ISOTROPIC T2-weighted: voxel size $0.84 \times 0.84 \times 0.94$ mm^3 to $0.84 \times 0.875 \times 0.875 \times 0.9$ mm^3

[*]Sex not reported for 7 patients
[†]Eight patients were followed up with more than 1 MRI examination
[‡]Five scans were acquired very late after injury, resulting in a high average time for MRI examination

scans (site 1) and 64 patients with T2-weighted axial scans (site 2). We used the region-based training strategy of nnUNet [173], where the model initially segments the spinal cord and then localizes itself on the spinal cord to segment the T2-weighted lesions subsequently. The lesions are segmented as a single object covering hyperintense and hypointense voxels, hence containing both edema and hemorrhage. The following default nnUNet data augmentation methods were used: random rotation, scaling, mirroring, Gaussian noise addition, Gaussian blurring, adjusting image brightness and contrast, low-resolution simulation, and Gamma transformation. All scans were preprocessed in right, posterior, inferior (RPI) orientation, resampled to a common resolution ($0.78 \times 0.56 \times 0.78$ mm^3, corresponding to the median of all

image resolutions in the training set) and intensity-normalized using Z-score normalization. The model was trained for 1000 epochs, with a batch size of 2 using the stochastic gradient descent optimizer with a polynomial learning rate scheduler.

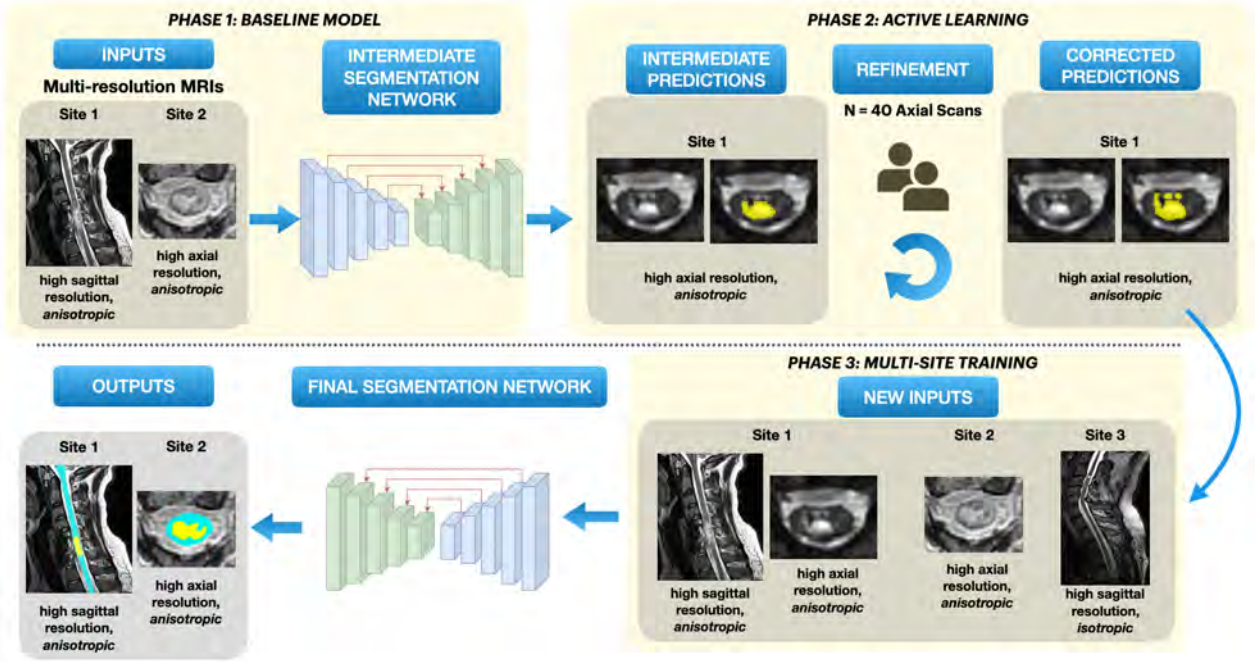

Figure 4.2 **Overview of our segmentation approach**. Phase 1: A baseline model is trained on data consisting of T2-weighted scans with axial and sagittal orientations from two sites. Phase 2: Active learning – Initial batch of automatic predictions on T2-weighted axial scans from site 2 are obtained, followed by manual corrections. Phase 3: Along with the newly corrected axial scans, isotropic T2-weighted sagittal scans from site 3 are added to the original dataset for multi-site training. The final model is trained to segment both spinal cord and lesion simultaneously.

For the second phase, we used the human-in-the-loop active learning strategy [57] to gradually include axial T2-weighted scans from site 1 in the training dataset. Using the phase 1 baseline model, we generated initial spinal cord and lesion predictions for unlabeled axial scans from site 1. In a subset of 40 scans, predicted segmentations underwent quality control, with two raters (E.N.K. and J.V.) manually correcting if needed. These refined segmentations were then added to the training dataset, resulting in 182 scans in the training set. For the third training phase, to further improve our model's generalization capabilities to a wide range of image resolutions, we added a new dataset from site 3 containing 14 isotropic resolution T2-weighted sagittal scans of patients with traumatic SCI.

In summary, the final dataset consisting of 196 scans from three sites was used for training the model with the region-based strategy described above.

4.2.4 Evaluation protocol

We created two independent test sets (site 1: $n = 19$, site 2: $n = 16$), following the 80/20 train/test splitting ratio (Table 4.1). To ensure an unbiased assessment of the model's performance and avoid overfitting, the train/test splits were done at the patient-level rather than the image-level, meaning that for the subset of 40 patients with sagittal and axial scans, both scans were included in the training split. For the test set, however, only the sagittal scans of each patient were used. We trained five models, each with a different train/test splitting using a different random seed, to avoid biasing the model towards a particular dataset split. We chose to use 5 random seeds instead of one 5-fold cross-validation to increase the likelihood of more diverse test sets containing SCI patients with (i) well-defined hyperintense lesions and (ii) lesions appearing under severe metal implants. The model's performance in spinal cord segmentation was evaluated independently within each test set by comparing it with open-source methods available in Spinal Cord Toolbox (SCT) [86] : `sct_propseg` [174], `sct_deepseg_sc` [17], and the recently proposed contrast-agnostic spinal cord segmentation model [175]. Due to the lack of existing state-of-the-art, open-source methods for SCI lesion segmentation, we compared the SCIseg 3D model with its 2D version. Additionally, we tested our model on an independent cohort of 14 patients with DCM from site 1 (unseen during training) to evaluate its generalization on patients with non-traumatic SCI (Figure 4.1).

4.2.5 Evaluation metrics

For quantitative evaluation of model performance, we used the segmentation metrics from the open-source ANIMA toolkit https://anima.readthedocs.io/en/latest/index.html. For spinal cord segmentation, we calculated the Dice score and the relative volume error (RVE). For lesion segmentation, we calculated the Dice score, average surface distance, lesion-wise positive predictive value (PPV), lesion-wise sensitivity, and F1 score [176]. As we trained five models on five random train/test splits (instead of 5-fold cross-validation), some patients were present in more than one test set. We thus averaged the metrics across test splits for such patients.

4.2.6 Quantitative MRI biomarkers

We used the SCT's `sct_analyze_lesion` function to automatically compute the total lesion volume, intramedullary lesion length, and maximal axial damage ratio [103] from the manual reference standard lesion masks and the automatic predictions using the proposed SCIseg 3D model. To assess the effect of adding more training data during active learning, we computed

the quantitative MRI biomarkers before (phase 1) and after active learning (phase 3). The quantitative MRI biomarkers were then averaged across five random test splits. Additionally, for site 2, we correlated the quantitative MRI biomarkers with the clinical scores (light touch, pinprick, and lower extremity motor scores).

4.2.7 Statistical analysis

Statistical analysis was performed using the SciPy Python library v1.11.4 [177]. Data normality was tested using the D'Agostino and Pearson's normality test. Within-site comparisons of age and sex between patients from the testing and training sets were performed using the Mann-Whitney U test and the chi-squared test, respectively. Between-group comparisons (spinal cord segmentation performance SCIseg vs. open-source methods; lesion segmentation performance SCIseg 2D vs. SCIseg 3D; SCIseg lesion segmentation performance before (phase 1) vs. after active learning (phase 3); manual reference standard lesion masks vs. SCIseg-predicted lesions) were performed using the Wilcoxon signed-rank test. The SCIseg lesion segmentation performance between sites 1 and 2 was tested using the Mann-Whitney U test. Correlations between clinical scores and quantitative MRI biomarkers were examined using the Spearman rank-order correlation. $P < .05$ was considered to indicate a statistically significant difference.

4.3 Results

4.3.1 Patient characteristics

A total of 191 patients (mean age ± standard deviation 48.1 ± 17.9 years; 142 males, 42 females, 7 sex not reported) with 231 MRI scans from three sites with different lesion etiologies (traumatic, ischemic, and hemorrhagic) were included in this study (Figure 4.1). Ninety-seven patients were from site 1, of which 61 had surgical hardware (dorsal or ventral spondylodesis), 13 underwent decompressive surgery, and the remaining 23 patients did not undergo surgery. Eighty patients were from site 2, all of which had postoperative metallic stabilization. Lastly, 14 patients from site 3, of which 8 had surgical hardware, 2 had decompressive surgery, and 4 patients did not undergo any surgery. Details about patient demographics, injury levels, injury chronicity and scanner types can be found in Table 4.1. Eight patients from site 1 were followed up with additional MRI examinations. The final training set comprised 196 MRI scans, and the final testing set contained 35 MRI scans. There was no evidence of within-site differences in age (site 1: $P = .06$, site 2: $P = .20$) and sex (site 1: $P = .71$, site 2: $P = .72$) between patients from the testing and training

sets. Patients were scanned across scanners from different manufacturers (Siemens, Philips, GE) with different field strengths (1T, 1.5T, 3T). T2-weighted scans used in this study had heterogeneous image resolutions and orientations (Table 4.1).

4.3.2 Automatic spinal cord and lesion segmentation in SCI

Table 4.2 shows the quantitative results of SCIseg 3D on test sets of the two sites stratified by scanner strength along with their averages. As shown by the Dice scores, spinal cord segmentations from the model were stable across different data splits and magnetic field strengths despite the presence of MRI artifacts induced by spinal hardware in the scans. However, for lesion segmentation, the model performed better on site 2 compared to site 1 (Dice score for site 2: 0.74 ± 0.15 vs Dice score for site 1: 0.51 ± 0.30), with a high standard deviation across splits.

Table 4.2 Spinal cord and lesion segmentation performance of the proposed SCIseg 3D model on the test set.

	Spinal Cord Segmentation						
	Site 1 ($n = 79$) T2-weighted sagittal			Site 2 ($n = 51$) T2-weighted axial			Site 1 & Site 2 ($n = 130$)
Metric	1.5T ($n = 33$)	3T ($n = 46$)	1.5T and 3T ($n = 79$)	1.5T ($n = 36$)	3T ($n = 15$)	1.5T and 3T ($n = 51$)	1.5T and 3T ($n = 130$)
Dice Score (↑)	0.88 ± 0.08	0.92 ± 0.08	0.90 ± 0.08	0.94 ± 0.04	0.94 ± 0.02	0.94 ± 0.04	0.92 ± 0.07
RVE % (↓)	0.35 ± 22.32	0.17 ± 6.76	0.25 ± 15.19	1.28 ± 11.58	-1.34 ± 6.39	0.51 ± 10.33	0.35 ± 13.45
Surface Distance (↓)	0.25 ± 1.00	0.06 ± 0.18	0.14 ± 0.66	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.09 ± 0.52
	Lesion Segmentation						
Dice Score (↑)	0.47 ± 0.30	0.54 ± 0.30	± 0.30	0.78 ± 0.10	0.67 ± 0.23	0.74 ± 0.15	0.61 ± 0.27
Surface Distance (↓)	5.25 ± 13.43	2.28 ± 5.43	3.51 ± 9.61	0.06 ± 0.14	0.91 ± 2.48	0.30 ± 1.34	2.17 ± 7.54
Lesion-wise PPV % (↑)	51 ± 41	54 ± 45	53 ± 43	91 ± 19	81 ± 37	88 ± 26	67 ± 41
Lesion-wise Sensitivity % (↑)	76 ± 38	80 ± 35	79 ± 36	90 ± 20	93 ± 26	91 ± 22	84 ± 32
F_1 Score (↑)	0.54 ± 0.41	0.55 ± 0.44	0.55 ± 0.43	0.88 ± 0.18	0.82 ± 0.36	0.86 ± 0.24	0.68 ± 0.39

Note: The metrics are averaged across 5 random seeds and reported as means ± standard deviations. The best values for surface distance and RVE are 0.0 and 1.0 for the rest of the metrics. Up arrows indicate that the higher the values the better and vice-versa for the down arrows.
RVE = relative volume error, PPV = positive predictive value.

4.3.3 Comparison with other methods

The comparison of spinal cord and SCI lesion segmentation performance of SCIseg 3D with other methods is shown in Figures 4.3 and 4.4. The half-violin plots in Figure 4.4 show the

distribution of the Dice scores and RVE for test scans across all seeds and the scatter plots show the performance of the models on each test scan.

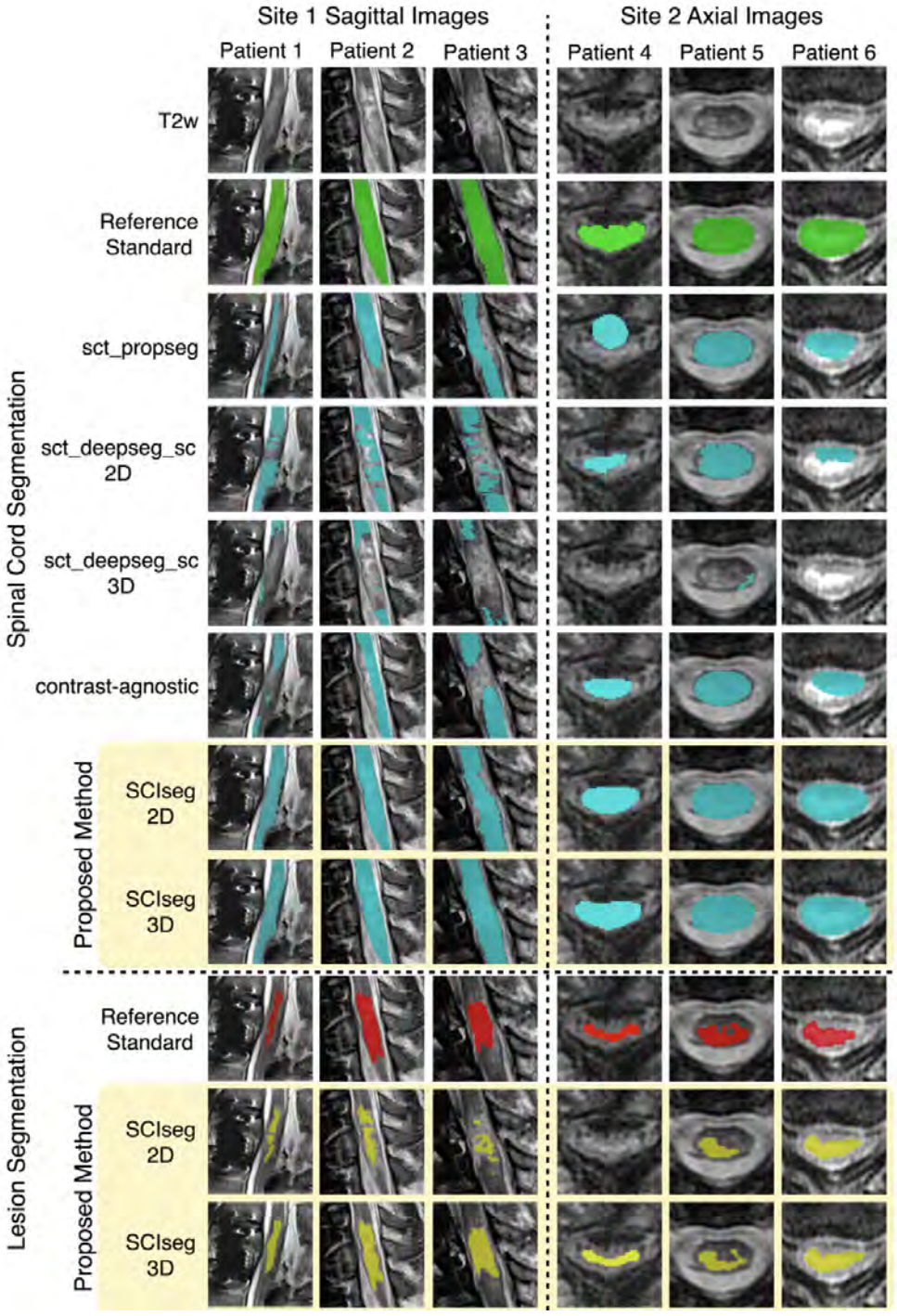

Figure 4.3 Comparison of SCIseg with baseline methods for the spinal cord and lesion segmentation on patients from site 1 and site 2. SCIseg 3D provides the best qualitative results for both spinal cord and lesion segmentation. T2w = T2-weighted

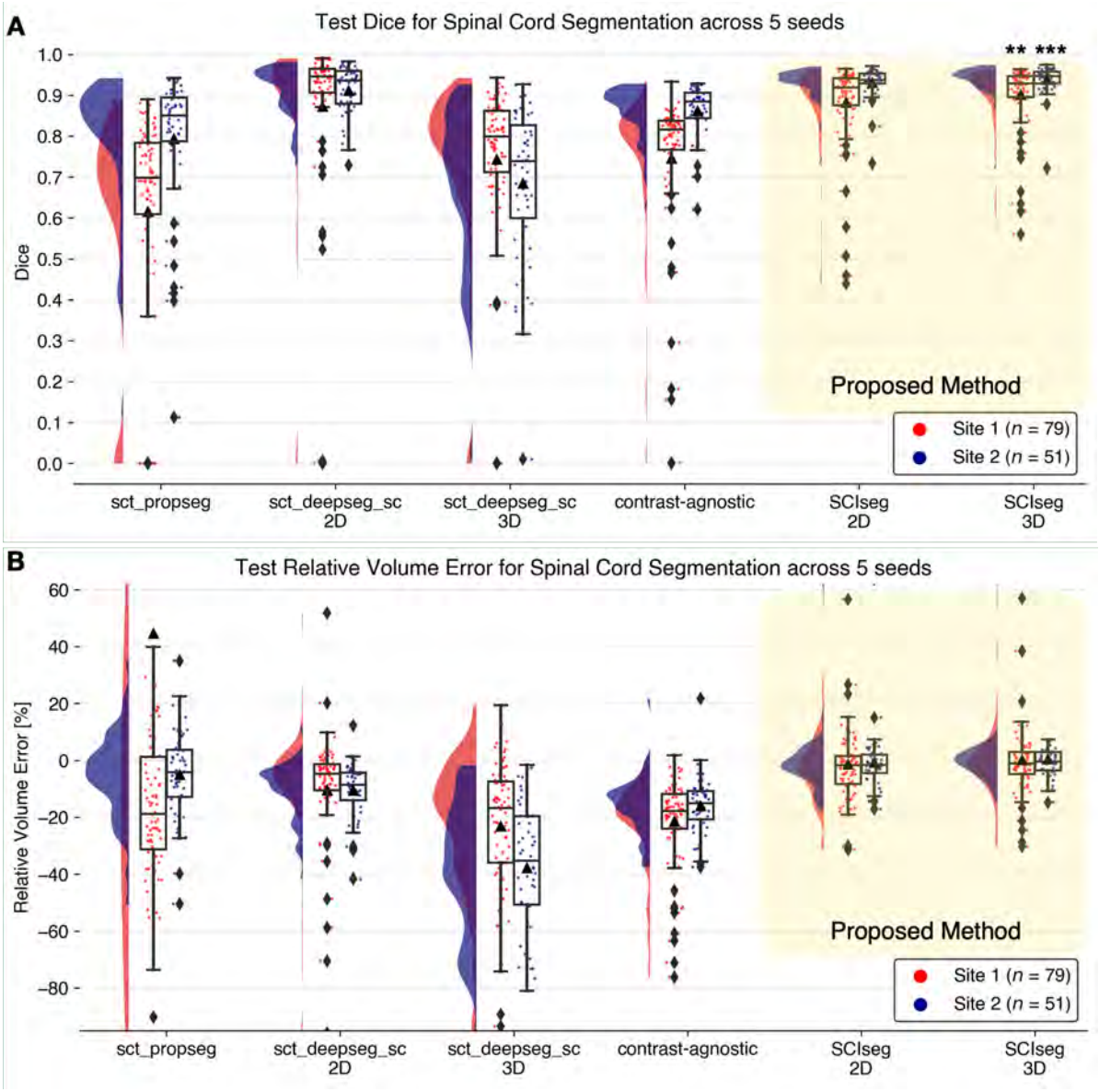

Figure 4.4 Raincloud plots comparing the (A) Dice scores (best: 1; worst: 0) and (B) relative volume error (in %, best: 0%) across various spinal cord segmentation methods. The numbers in the legend represent the number of test scans in each site across 5 different training seeds. Notice that although the `sct_deepseg_sc` 2D and `SCIseg` 3D have similar Dice scores, the former shows a higher under-segmentation (negative relative volume error) compared to the latter. $***P < .05$ (two-sided Bonferroni-corrected pairwise Wilcoxon signed-rank test for SCIseg 3D with all baselines), $**P < .001$ (statistically significant for all pairs except `SCIseg` 3D and `sct_deepseg_sc` 2D).

Spinal cord segmentation

SCIseg 3D achieved the best segmentation performance (Dice score of 0.92 ± 0.07; RVE of 0.35 ± 13.45) for both sites compared to other baselines (Figure 4.4). For site 1, the

Wilcoxon signed-rank test revealed significant differences ($P < .001$, Bonferroni-corrected) in Dice score between SCIseg 3D and all baseline methods except for `sct_deepseg_sc` 2D. As described in Section 2.2, this is the consequence of using a semi-automatic approach involving `sct_deepseg_sc` 2D to create the reference standard spinal cord segmentation masks. As a result, quantitative evaluations involving the reference standard masks obtained from this baseline model are inherently biased to be higher than the rest of the methods in comparison. Although there was no significant difference between SCIseg 3D and `sct_deepseg_sc` 2D, the SCIseg 3D model obtained spinal cord segmentations for all test cases, including those where the baselines obtained empty predictions (shown by diamonds at Dice=0 in Figure 4.4A). For site 2, statistically significant differences in Dice scores were found between SCIseg 3D and all baselines ($P < .05$, Bonferroni-corrected). We observed more under- and over-segmented predictions by SCIseg 3D for site 1 relative to site 2 (shown by a larger spread of scatter points around RVE=0% in Figure 4.4B). On visual quality control of such cases, we found that the scans contained substantial metal implants, interfering with the model's ability to fully segment the spinal cord. Lastly, it must be noted that all baselines were trained specifically for segmenting the spinal cord, whereas the proposed SCIseg 3D model can segment both the spinal cord and SCI lesions simultaneously.

Lesion segmentation

Table 4.3 shows a comparison between the 2D and 3D variants of the SCIseg model. The 3D model performed significantly better than the 2D model for both sites. As for the performance between sites, the 3D model's performance on site 2 was higher than that of site 1 (see Table 4.3 for the p-values). Through visual quality control, we found that site 1 contained several patients with metal implants causing heavy image artifacts. Additionally, SCI lesions spanned different phases (acute and sub-acute) with various degrees of lesion hyperintensity, thus making automatic segmentation challenging.

4.3.4 Effect of active learning on lesion segmentation

We performed an ablation study comparing the model performance after phase 1 (training on 2 sites) and phase 3 (training on 3 sites after active learning). Figure 4.5A shows the correlation between manual GT and automatic predictions for total lesion volume (top) and intramedullary lesion length (bottom). For both sites, higher agreement between the manually annotated and automatically derived lesion metrics was observed for the final model after the third phase of training (i.e., solid lines moving closer to the diagonal identity line). There was a statistically significant improvement after active learning (phase 3) in estimating

Table 4.3 Comparison of lesion segmentation performance between SCIseg 2D and 3D models.

	SCIseg 2D		**SCIseg 3D**	
Metric	Site 1 ($n = 79$)	Site 2 ($n = 51$)	Site 1 ($n = 79$)	Site 2 ($n = 51$)
Dice Score (↑)	0.36 ± 0.31	0.65 ± 0.21*	**0.51 ± 0.30**†	**0.74 ± 0.15***
Surface Distance (↓)	9.14 ± 32.43	0.54 ± 1.48*	**3.51 ± 9.61**‡	**0.30 ± 1.34****
Lesion-wise PPV % (↑)	33 ± 43	73 ± 35*	**53 ± 43**†	**88 ± 26**$^{\ddagger *}$
Lesion-wise Sensitivity % (↑)	63 ± 46	89 ± 27**	**79 ± 36**†	**91 ± 22**
F_1 Score (↑)	0.35 ± 0.43	0.74 ± 0.33*	**0.55 ± 0.43**†	**0.86 ± 0.24** $^{\ddagger *}$

Note: The metrics are averaged across 5 different training seeds and reported as means ± standard deviations. The best value for surface distance is 0.0 and 1.0 for the rest of the metrics. Up arrows indicate that the higher the values the better and vice-versa for the down arrows.
RVE = relative volume error, PPV = positive predictive value
†Statistically significant compared to SCIseg 2D. Wilcoxon signed-rank test ($P < 0.001$)
‡Statistically significant compared to SCIseg 2D. Wilcoxon signed-rank test ($P < 0.05$)
*Statistically significant compared to Site 1. Mann-Whitney U test ($P < 0.001$)
**Statistically significant compared to Site 1. Mann-Whitney U test ($P < 0.05$)

the total lesion volume ($P = .016$) and the lesion length ($P = .004$) for site 1.

Figure 4.5B shows the lesion segmentation performance of our baseline model after phase 1 of training (before active learning) on unseen axial T2-weighted scans from site 1. Qualitatively, the model tends to under-segment the lesions. However, there was an overall improvement in segmentation performance when the model was trained on more data consisting of axial scans from site 1 and isotropic sagittal scans from site 3 during phase 3 of training.

4.3.5 Generalization to degenerative cervical myelopathy

Qualitative examples of spinal cord and lesion segmentation on an independent dataset of patients with DCM whose data were unseen during model training are shown in Figure 4.5C. Interestingly, in cases where the reference standard lesion masks were under-segmented, the model provided a better and more complete segmentation of the lesion. Furthermore, the spinal cord segmentations were accurate even for slices with severe spinal cord compression (Figure 4.5C, Patient 12). Table 4.4 shows the Dice and F1 scores for both spinal cord and lesion segmentations. Despite not being trained on DCM lesions, the model achieved a high Dice score of 0.95 for spinal cord segmentation and an F1 score of 0.49 for lesion segmentation.

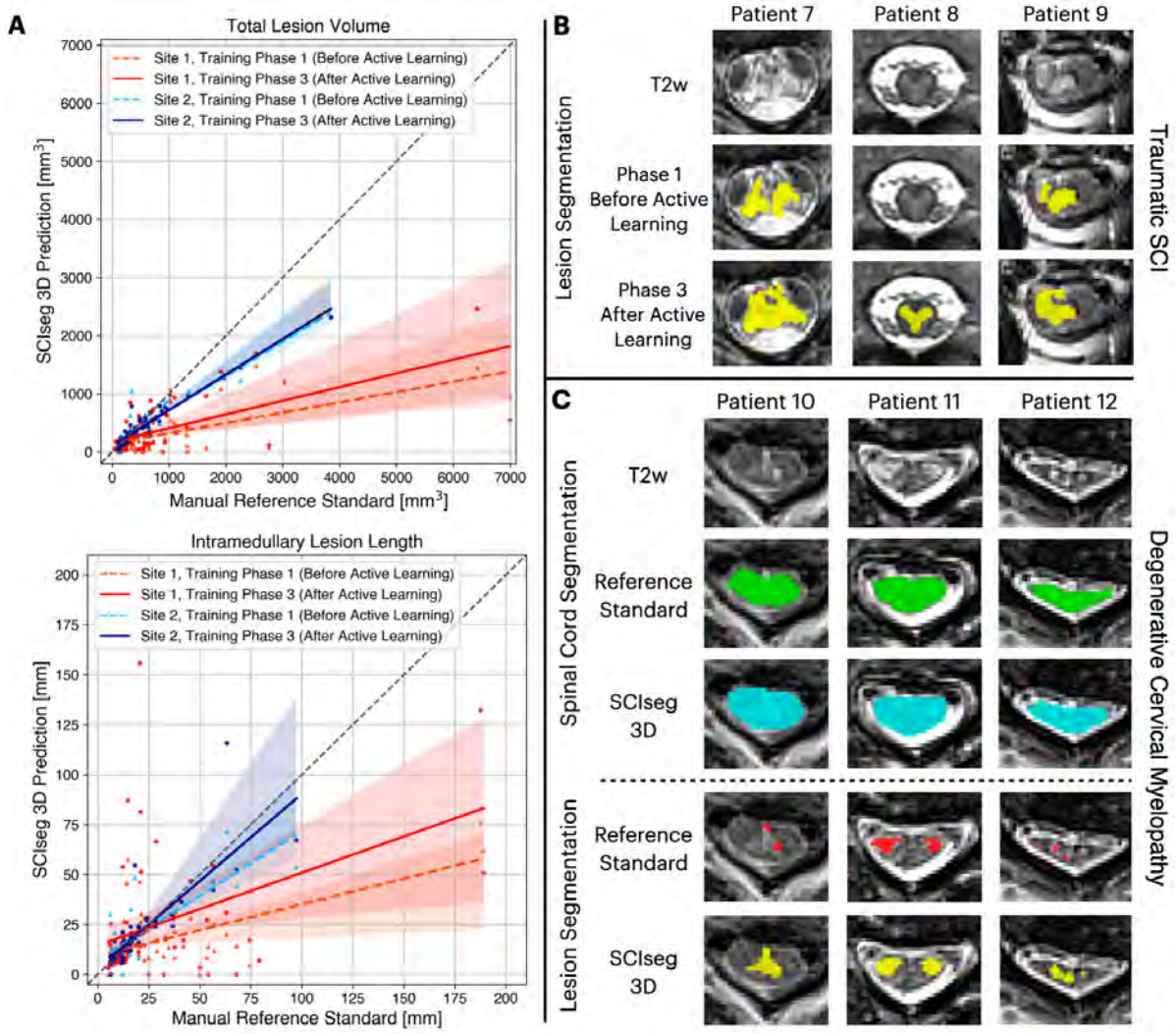

Figure 4.5 Comparison of model performance before and after active learning. (A) Correlation plots for total lesion volume (top) and intramedullary lesion length (bottom) computed from the manual reference standard lesion masks (x-axis) and lesion segmentation predictions from the proposed SCIseg 3D model (y-axis). Within each plot, colored dashed and solid lines show the agreement between the manual reference standard and automatic predictions before and after active learning, respectively, for site 1 (red/orange) and site 2 (blue/light-blue). Note that the model's predictions after active learning show a higher agreement with the manual reference standard for both sites (i.e., solid lines move closer to the diagonal identity line). (B) SCIseg's predictions on unseen axial scans from site 2 before and after active learning. (C) Examples of SCIseg's generalization to patients with non-traumatic SCI (i.e., degenerative cervical myelopathy, DCM). Notice that the model obtains an accurate spinal cord segmentation even at the level of severe compression (Patient 12). T2w = T2-weighted.

Table 4.4 Quantitative evaluation of generalizability of the SCIseg 3D model to patients with non-traumatic SCI (i.e., DCM).

Metric	**Spinal Cord Segmentation**
Dice Score (↑)	0.95 ± 0.01
RVE % (↓)	3.51 ± 4.63
Surface Distance (↓)	0.0 ± 0.0
	Lesion Segmentation
Dice Score (↑)	0.46 ± 0.26
Surface Distance (↓)	1.51 ± 4.64
Lesion-wise PPV % (↑)	71 ± 38
Lesion-wise Sensitivity % (↑)	50 ± 45
F_1 Score (↑)	0.49 ± 0.44

Note: The metrics are averaged across 5 random seeds and reported as means ± standard deviations. The best value for surface distance and RVE is 0.0 and 1.0 for the rest of the metrics. Up arrows indicate that the higher the values the better and vice-versa for the down arrows.
DCM = degenerative cervical myelopathy, RVE = relative volume error, PPV = positive predictive value, SCI = spinal cord injury.

4.3.6 Manual vs. SCIseg-predicted lesion biomarkers

Quantitative MRI biomarkers obtained from SCIseg 3D predictions were comparable with those obtained from manually segmented lesions (Figure 4.6). The Wilcoxon signed-rank test between SCIseg-predicted (Figure 4.6, green) vs. manual (Figure 4.6, yellow) lesion biomarkers revealed a significant difference in lesion volume ($P = .003$) between the two groups and no evidence of a difference for lesion length ($P = .42$) and maximal axial damage ratio ($P = .16$).

4.3.7 Correlation between clinical scores and MRI biomarkers

Figure 4.6 shows the correlation plots (including correlation coefficients and p-values) between clinical scores and quantitative MRI biomarkers calculated from both manual reference standard lesion masks and lesions segmented using SCIseg 3D.

4.4 Discussion and Conclusion

This study introduced a DL-based model, SCIseg, for the automatic segmentation of the spinal cord and intramedullary lesions in patients with SCI on T2-weighted MRI scans. The model was trained and evaluated on a cohort of 191 patients with traumatic and non-traumatic SCI with 231 scans acquired using different scanner manufacturers with heteroge-

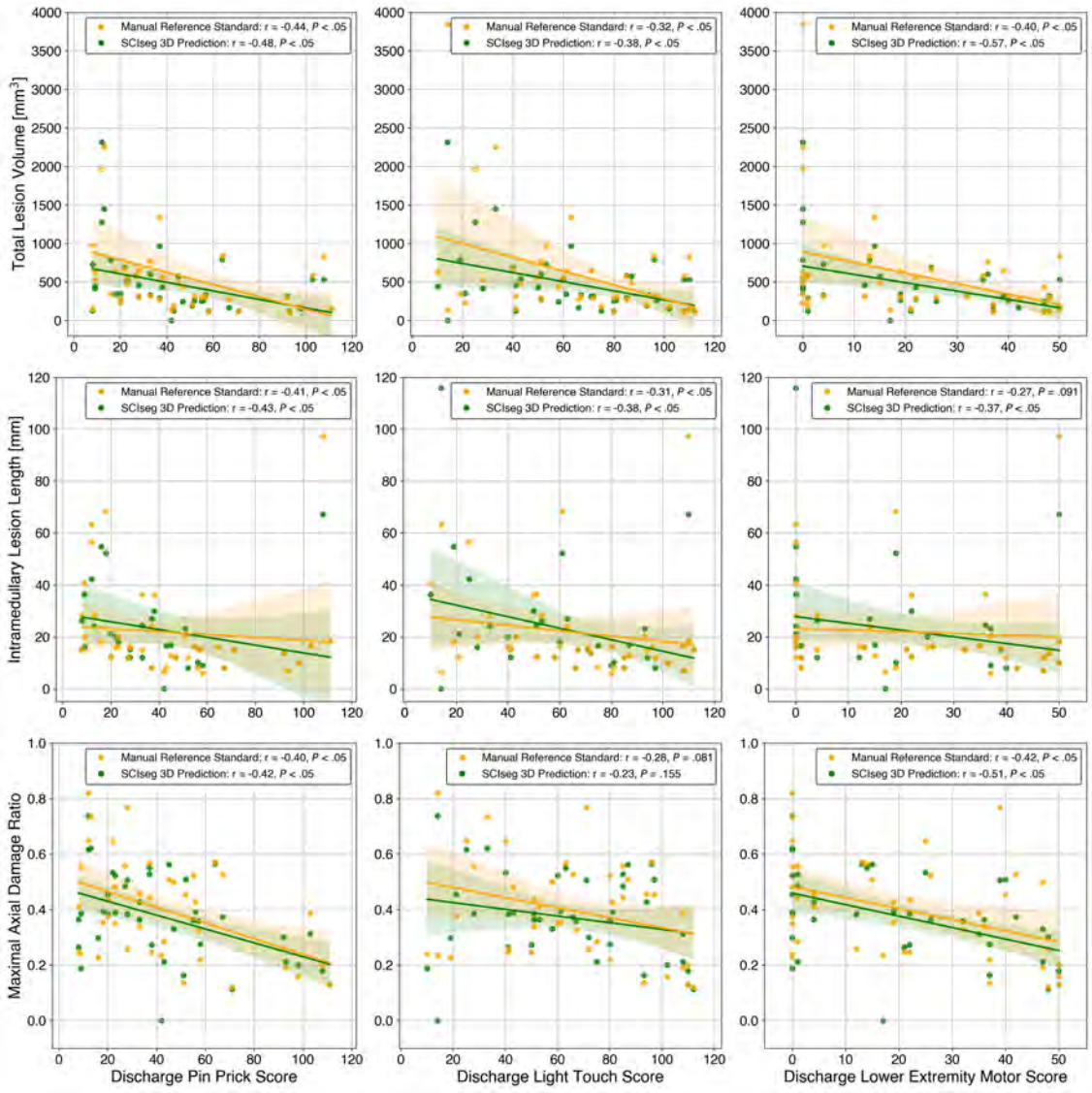

Figure 4.6 Correlation analysis between discharge clinical scores (x-axis) and quantitative MRI biomarkers (y-axis) for site 2. Spearman correlation coefficient and p-value are shown in the legends of individual subplots. The Wilcoxon signed-rank test between the manual reference standard lesion masks (yellow) vs. automatic predictions using SCIseg 3D (green) lesion biomarkers revealed no evidence of differences for lesion length ($P = .42$) and maximal axial damage ratio ($P = .16$), but a significant difference for lesion volume ($P = .003$).

neous image resolutions (isotropic and anisotropic), and orientations (axial and sagittal). Patients had various lesion etiologies (traumatic, ischemic, and hemorrhagic) and lesions spread across the cervical, thoracic and lumbar spine. SCIseg achieved a Dice score of 0.92 ± 0.07 (mean $\pm$ SD) for spinal cord segmentation and 0.61 ± 0.27 for SCI lesion segmentation. There was no evidence of a difference between lesion length ($P = .42$) and maximal axial damage ratio ($P = .16$) computed from manually annotated lesions and SCIseg predictions.

To the best of our knowledge, SCIseg is the first open-source, automatic method for lesion and spinal cord segmentation in SCI. It also generalizes to DCM patients, producing accurate segmentations for both lesions and the spinal cord at the compression levels.

As the segmentation performance might be constrained by the low data quality and small dataset sizes in SCI, we showed that implementing a three-phase training strategy, including an active learning approach to progressively expand the dataset size and incorporating diverse data distributions into the training set, enhances the model's performance. Furthermore, a region-based training strategy that jointly segments the spinal cord and the lesion is more efficient than training two individual models for spinal cord and lesion segmentation, respectively. As a result, correlation analyses between clinical scores and MRI-derived biomarkers showed statistically significant relationships for both manually annotated reference standard and automatically derived lesion masks, suggesting the SCIseg predictions can be reliably used for correlation with clinical measures in SCI.

Our cohort predominantly consisted of traumatic SCI lesions in intermediate and chronic phases, as the prevalence of ischemic and hemorrhagic lesions is typically lower [154]. As chronic injuries tend to be more delineated on T2w scans [34], our model learned to be sensitive to hyperintense abnormalities in the image. This also explains its ability to segment DCM lesions, which also tend to be hyperintense at the site of compression. Similarly, as the injury levels in the training dataset were skewed towards the cervical spine, the model's ability to segment lumbar lesions is expected to be lower when compared to cervical or thoracic lesions. Despite the presence of metal implants causing strong image artifacts in several patients, SCIseg provides a good starting point for obtaining lesion and spinal cord segmentations instead of manual annotations from scratch.

Only a few studies exist in the literature discussing the importance of automatic segmentation in SCI scans [168, 178]. The study by McCoy et al. [168] is most similar to ours, as it presented the first DL method for segmentation of the spinal cord and intramedullary lesions in SCI. Nevertheless, there are several important distinctions between the two studies. While their model was trained on axial preoperative scans of patients with acute SCI from a single site, our model was trained on multi-site data consisting of patients with traumatic, ischemic, and hemorrhagic SCI with different image orientations (axial and sagittal). Moreover, our model was exposed to more heterogeneous data covering different injury phases (intermediate and chronic) and therefore demonstrated better generalization to both traumatic and non-traumatic lesion etiologies. More importantly, our work is open-source, further enabling reproducible, multi-site studies in SCI. There exist several promising avenues for future work. The segmentation models can be improved by using more fine-grained reference

standard masks, where the hyperintense edema and hypointense hemorrhage could be treated as separate classes. Training a model on pre-operative traumatic SCI data using these reference standard masks would have a major impact on improving the initial classification of the disease and further prognostication [163]. While the model generalizes reasonably well to DCM lesions, there is potential for improvement, particularly by adding the DCM cohort to the existing training set or by training a DL model exclusively on DCM data. Previous studies have reported the presence of hyperintense T2-weighted lesions in up to 64% of patients with DCM [161, 179, 180] and explored the relationship between structural and functional damages [181]. Such studies would greatly benefit from an automatic DCM lesion segmentation method. Aggregating more heterogeneous data will also allow for sensitivity analysis of potential confounding factors.

This study had some limitations. First, longitudinal scans from patients with follow-up examinations were treated as independent inputs for training. While the lesion appearance evolved between sessions, resulting in non-identical lesions (hence justifying our choice of treating them as independent inputs), the model likely could not learn the evolution of lesions across time. Second, the model's sensitivity to hyperintense abnormalities might result in false positive segmentations in healthy controls where the spinal cord central canal is visualized. Third, our limited size of 196 MRI scans in the training set risks overfitting, given the complexity of the SCI lesion segmentation task. While we gathered diverse data from 3 sites and trained 5 models on different train/test splits, along with extensive data augmentation to prevent potential overfitting, increasing the dataset size would further improve the model's performance and generalization. Lastly, we did not analyze the inter-rater variability as the data were gathered from multiple sites and there were no overlapping patients across sites. Previous studies [182, 183] have reported that MRI measures of spinal cord damage (e.g., edema length, midsagittal tissue bridge ratio, axial damage ratio) do exhibit high-to-excellent levels of inter-rater reliability.

In conclusion, this study presented SCIseg, an automatic DL-based method for the segmentation of the spinal cord and intramedullary lesions in SCI on T2-weighted MRI scans. The work has addressed several limitations of previous studies by including a large retrospective cohort consisting of 191 patients spanning three sites, using MRI data acquired with scanners from different manufacturers, and training a single model to segment both the spinal cord and SCI lesions. More importantly, the methodology has been designed to ensure reproducibility and enable large-scale, reproducible prospective studies. The model is open-source and accessible via SCT (v6.2 and higher). We hope that SCIseg will benefit clinicians and patients by providing additional diagnostic and prognostic information, serving as a basis for further studies assessing optimal rehabilitation from a customized patient-based perspective.

4.5 Future Work: Automatic Measurements of Tissue Bridges

For a more clinically-oriented application of the SCIseg, we proposed an approach to automatically measure the width of the tissue bridges from the predicted lesion mask [184]. We refer the reader to Appendix A for more details.

CHAPTER 5 ARTICLE 2: TOWARDS CONTRAST-AGNOSTIC SOFT SEGMENTATION OF THE SPINAL CORD

Authors Sandrine Bédard[*1], Enamundram Naga Karthik[*,1,2], Charidimos Tsagkas[3,4,5], Emanuele Pravatà[6], Cristina Granziera[3,4,5], Andrew Smith[7], Kenneth Arnold Weber II[8], Julien Cohen-Adad[1,2,9,10]

Affiliations

1. NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montréal, Montréal, Québec, Canada
2. Mila - Québec Artificial Intelligence Institute, Montréal, Québec, Canada
3. Translational Imaging in Neurology (ThINK) Basel, Department of Biomedical Engineering, Faculty of Medicine, University Hospital Basel and University of Basel, Basel, Switzerland
4. Department of Neurology, University Hospital Basel, Basel, Switzerland
5. Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland
6. Neuroradiology Department, Neurocenter of Southern Switzerland, Ospedale Regionale di Lugano, Lugano, Switzerland
7. Department of Physical Medicine and Rehabilitation Physical Therapy Program, University of Colorado School of Medicine, Aurora, CO, USA
8. Division of Pain Medicine, Department of Anesthesiology, Perioperative, and Pain Medicine, Stanford University School of Medicine, Stanford, CA, USA
9. Functional Neuroimaging Unit, CRIUGM, University of Montreal, Montreal, Québec, Canada
10. Centre de recherche du CHU Sainte-Justine, Université de Montréal, Montréal, Québec, Canada

[*] Shared co-first authorship - authors contributed equally

This article has been published as an open-access article [18] at the *Medical Image Analysis* journal on 2025-01-13 as:

> Bédard, S., Karthik, E.N., Tsagkas, C., Pravatà, E., Granziera, C., Smith, A., Weber II, K.A. and Cohen-Adad, J., 2025. Towards contrast-agnostic soft segmentation of the spinal cord. Medical Image Analysis, p.103473..

Contributions

Naga Karthik (NK) and Sandrine Bédard (SB) co-led the project. SB curated the dataset and developed the preprocessing framework to generate soft ground truth masks. NK trained the segmentation models, ran experiments comparing the proposed approach with other methods and performed the ablation studies. Both SB and NK developed the script to compute cross-sectional area and wrote the draft of the paper. Charidimos Tsagkas, Emanuele Pravata, Cristian Granziera, Andrew Smith and Kenneth Weber were involved in data acquisition of MS and SCI patients. Julien Cohen-Adad provided supervision throughout the project. All co-authors reviewed the draft of the manuscript.

Highlights

1. Automatic method for spinal cord segmentation robust to various MRI contrasts.
2. Network outputs 'soft' segmentation that encodes partial volume information.
3. Reduces variability in morphometric measures (cross-sectional area) across contrasts.
4. Generalizes well to unseen datasets, contrasts, pathologies, and MRI vendors.
5. Available in Spinal Cord Toolbox (SCT, v6.2 and above).

An extension of this work has been published in the "Short Paper" track at the *Medical Imaging with Deep Learning* (MIDL) conference as an open-access article:

> Karthik, E.N., Bedard, S., Valosek, J., Chandar, S. and Cohen-Adad, J., 2024. Contrast-agnostic Spinal Cord Segmentation: A Comparative Study of ConvNets and Vision Transformers. In Medical Imaging with Deep Learning. https://openreview.net/forum?id=n6D25aqdV3

Abstract

Spinal cord segmentation is clinically relevant and is notably used to compute spinal cord cross-sectional area (CSA) for the diagnosis and monitoring of cord compression or neurodegenerative diseases such as multiple sclerosis. While several semi and automatic methods exist, one key limitation remains: the segmentation depends on the MRI contrast, resulting in different CSA across contrasts. This is partly due to the varying appearance of the boundary between the spinal cord and the cerebrospinal fluid that depends on the sequence and acquisition parameters. This contrast-sensitive CSA adds variability in multi-center studies where protocols can vary, reducing the sensitivity to detect subtle atrophies. Moreover, existing methods enhance the CSA variability by training one model per contrast, while also producing binary masks that do not account for partial volume effects. In this work, we present a deep learning-based method that produces soft segmentations of the spinal cord that are stable across MRI contrasts. Using the Spine Generic Public Database of healthy participants (n = 267; contrasts = 6), we first generated participant-wise soft ground truth (GT) by averaging the binary segmentations across all 6 contrasts. These soft GT, along with aggressive data augmentation and a regression-based loss function, were then used to train a U-Net model for spinal cord segmentation. We evaluated our model against state-of-the-art methods and performed ablation studies involving different GT mask types, loss functions, contrast-specific models and domain generalization methods. Our results show that using the soft average segmentations along with a regression loss function reduces CSA variability ($p < 0.05$, Wilcoxon signed-rank test). The proposed spinal cord segmentation model generalizes better than the state-of-the-art contrast-specific methods amongst unseen datasets, vendors, contrasts, and pathologies (compression, lesions), while accounting for partial volume effects.

Keywords Spinal Cord, MRI, Contrasts, Segmentation, Deep Learning, Soft Labels, Partial Volume Effect

5.1 Introduction

Spinal cord segmentation is clinically relevant, notably to compute cross-sectional area (CSA) for the diagnosis and monitoring of atrophy in multiple sclerosis (MS) [185], spinal cord injury (SCI) [186], and in characterizing spinal cord compression [187]. While several approaches for semi-automatic and automatic segmentation of the spinal cord have been introduced [17, 188, 189], they all suffer from the same limitation: the output segmentation depends on the MRI contrast and acquisition parameters of the input image [17]. For instance, the CSA measured from T2-weighted (T2w) images is approximately 8% higher than that measured from T1-weighted (T1w) images [16, 190]. This contrast-dependency is partly due to the varying appearance of the boundary between the spinal cord and the cerebrospinal fluid because of differences in MR properties (e.g., relaxation times, spin density, flow). Different acquisition parameters and pulse sequences produce different contrast and sharpness of the spinal cord boundary, which consequently affect the output of the segmentation methods, whether they are manual, semi-automatic or automatic. The contrast-sensitive CSA also adds variability in multi-site studies, thereby reducing the sensitivity to detect subtle atrophies [102].

One way to mitigate the impact of MRI contrast on metrics derived from the segmentation is to compute the CSA on various contrasts and estimate a scaling factor based on the contrast type as done by [16] for T1w and T2w contrasts and MRI vendors. However, the scaling factors themselves are highly dependent on the MRI vendor, pulse sequence and imaging parameters, thus limiting their application to other studies.

Recent work addressed the contrast dependency of automatic segmentation in terms of model performance. Gros *et al.* [17] trained separate deep learning models for each contrast. However, the ground truth (GT) masks used for training were generated using a combination of automatic PropSeg [174] and manual corrections, which were done separately for each contrast. As a result, the GT masks for each contrast were already biased, resulting in models that robustly segmented the spinal cord, but produced different CSA across contrasts (e.g., higher T2w CSA than T1w CSA). SynthSeg [51, 69, 191], a deep learning-based method primarily used for the segmentation of brain MRI scans of any contrast or resolution, leveraged GT label maps during training to synthetically generate brain images of various contrasts. While SynthSeg provides segmentations that are inherently agnostic to the input image contrast, it relies on a domain randomization strategy, where parameters such as orientation, resolution, and contrast are randomly sampled from a uniform distribution to synthetically generate the training scans. However, this requires a large number of GT segmentations, which are difficult to obtain for spinal cord scans that often include various structures such

as the vertebrae, the spinal canal, intervertebral discs, nerve rootlets, surrounding muscles, the lungs and the heart.

Generalization to unseen domains is a paramount objective for deep learning algorithms. Domains can be defined as sets of images acquired from different sites and scanners, images consisting of contrasts other than those in training, or even images containing pathologies (i.e., lesions) when trained on healthy images. Domain generalization methods in the literature treat this as a domain shift problem at the fundamental level, where each contrast, for example, is seen as a different but related domain with minor differences in their marginal distributions [192]. Such methods propose to use domain adaptation techniques to *transfer* the differences between the source and target domains, either by mapping both domains to a shared latent space [193–195] or by generatively adapting the source to target domains by image-to-image translation methods [196, 197]. Expanding further into the concept of learning domain-invariant features, regularization, as a means of creating a representative feature space consisting of various domains has also been explored [63, 198, 199]. Other related works include meta-learning for adapting model for few-shot weakly-supervised segmentation tasks [200] and adversarial training for increasing the diversity of the training data [201]. While unsupervised methods in domain adaptation alleviate the need for labeled training data, such methods still need re-training on each subsequent target domain [202], which is impractical.

Data augmentation-based domain generalization methods aim to model the domain shifts via a series of transformations applied to the input images at the source domains during training. For instance, Zhao *et al.* [203] proposed a learning model for spatial transformations to synthesize additional labeled examples for one-shot segmentation in brain MRI scans. Ouyang *et al.* [204] used causality-driven data augmentation specifically targeting domain shifts and acquisition shifts, while Su *et al.* [68] proposed a location-scale augmentation using Bezier transformations, both in the context of single-source domain generalization. Ling *et al.* [67] showed that simply relying on sequential stack of data augmentation transforms based on image quality, appearance and spatial configuration, results in good generalization to unseen domains.

In addition to the contrast-dependent issues discussed above, one of the main limitations of traditional segmentation methods is that they produce binary (hard) segmentation masks, which do not account for partial volume effects [51, 205]. Partial volume effect is characterized by mixing of signals from different tissues within the same voxel, resulting in averaged intensities which are not representative of any of the underlying tissues. Binary masks do not provide calibrated output probabilities for the partial volume information of the tis-

sue. With soft labels, the segmentation is encoded with continuous values between 0 and 1 and can therefore encode partial volume information, while resulting in better generalization [15], faster learning [206], and increase the precision of voxel-based morphometry or CSA measurements [53].

5.1.1 Contributions

In this work, we present a convolutional neural network (CNN) model for the automatic soft segmentation of spinal cord across various contrasts. Our model reduces the variability in CSA across contrasts and generalizes to spinal cord images of unseen contrasts and pathologies. Our original contributions are as follows:

1. We introduce a new pipeline for generating a unique, soft GT that represents the segmentations across various MRI contrasts.
2. Contrary to [15] where the softness was obtained implicitly after data augmentation, we propose to apply the data-augmentation transforms directly on the soft GT masks and train a contrast-agnostic SoftSeg model for spinal cord segmentation.
3. We show that the proposed model reduces variability morphometric measures (i.e., produces stable soft segmentations across contrasts) and shows significant improvement over prior work using contrast-specific models and domain-generalization methods.

The model is open-source and the code for pre-processing/training/inference can be found in the following GitHub release[1]. It is also integrated into the Spinal Cord Toolbox and available in v6.2[2] and higher.

The rest of the paper is structured as follows: in Section 5.2, we describe the training dataset, preprocessing pipeline, the training and evaluation protocols. In Section 5.3, we show the results from the various validation experiments and comparisons with baselines and state-of-the-art methods. Different features and perspectives of the proposed spinal cord segmentation model are discussed in Section 5.4, followed by the conclusion.

[1]https://github.com/sct-pipeline/contrast-agnostic-softseg-spinalcord/tree/v2.0
[2]https://github.com/spinalcordtoolbox/spinalcordtoolbox/releases/tag/6.2

5.2 Materials and Methods

5.2.1 Dataset

We used the Spine Generic Public Database[3] (Multi-Subject) [16] consisting of 267 healthy participants scanned across multiple MRI vendors (Siemens, GE and Philips) and scanner models. Each participant has a 3D T1-weighted MPRAGE (T1w) at 1mm isotropic resolution, 3D T2w at 0.8 mm isotropic resolution, 2D T2*w axial at $0.5 \times 0.5 \times 3$ mm (multi-echo GRE), 3D axial gradient-echo with (MT-on) and without (GRE-T1w, with a shorter TR and a higher flip angle compared to the MT-on scan) magnetization transfer pulse at $0.9 \times 0.9 \times 5$ mm, and an axial diffusion-weighted scan motion corrected and averaged across diffusion directions at $0.9 \times 0.9 \times 5$ mm. This multi-contrast dataset was chosen because it is publicly available, and it includes a large variety of MR contrasts that are popular in the MR community. Participants with missing contrasts or excessive artifacts were excluded from our experiments ($n = 24$ out of 267).

The final dataset included 243 participants with 6 contrasts each, resulting in 1458 3D volumes in total. These were split according to 60/20/20 train/validation/test splits, resulting in 145 participants (870 volumes) for training, 49 participants (294 volumes) for validation and 49 participants (294 volumes) for testing.

5.2.2 Data preprocessing for ground truth generation

To eliminate the differences in CSA within the GT across contrasts, we used a unique segmentation averaged over all contrasts as the GT for training. Our objective here was to obtain the soft segmentation resulting from each contrast-specific hard segmentation. Figure 6.1 shows an overview of the procedure for generating the GT using SCT [86]. The GT soft segmentations are generated by averaging 6 different contrasts (T1w, T2w, T2*w, MT-on, GRE-T1w and DWI). For each participant and contrast, the spinal cord was segmented using SCT's `sct_deepseg_sc` to generate a binary segmentation. Manual corrections were made when significant segmentation errors (i.e., leaking and under-segmentation) were observed in SCT's quality control report. Since `sct_deepsg_sc` (DeepSeg2D) is considered the state-of-the-art for spinal cord segmentation and creating GT from scratch is highly time consuming, we obtained an initial batch of segmentations followed by manual quality control.

All images and binary segmentations were registered to the T2w image space as it has the highest resolution (0.8 mm isotropic). The registration was done using SCT's registration tool

[3] https://github.com/spine-generic/data-multi-subject

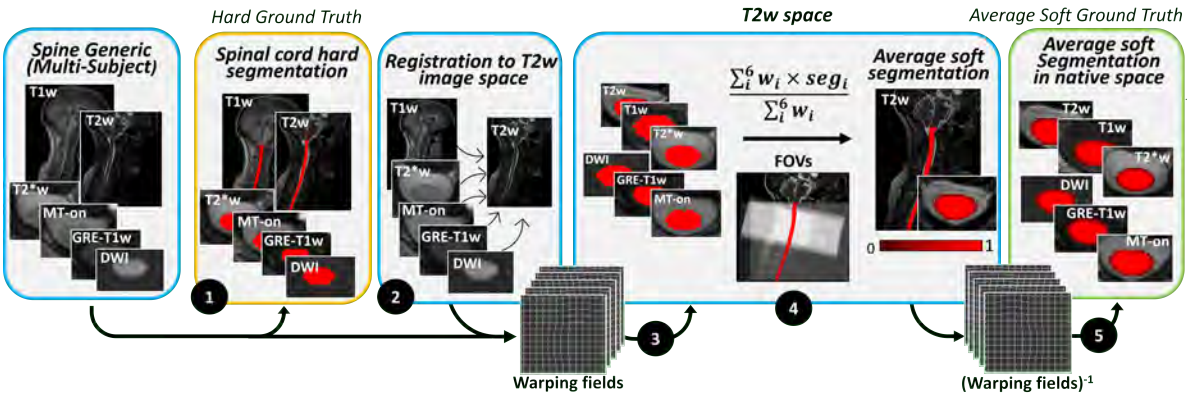

Figure 5.1 Preprocessing pipeline for soft average segmentations ground truth. (1) Automatic hard spinal cord segmentation using `sct_deepseg_sc` & manual corrections; (2) Registration to T2w space; (3) Applying each contrast's warping field to bring the segmentation masks to the T2w space; (4) Weighted averaging of segmentations according to each contrast FOV (represented by white rectangles) to create a unique soft GT mask (5) Applying inverse warping fields to bring the unique soft GT to the native space of each contrast.

`sct_register_multimodal` center-of-mass algorithm. It consists of a slice-by-slice alignment of the center of mass of the input and target segmentations (rotation and translation in x and y directions). The registration was performed in 10 iterations with a gradient step of 0.5. The segmentations of all 6 contrasts were then averaged within the T2w space to obtain a unique average soft segmentation (ranging from 0 to 1). The average of the segmentations was weighted according to the field-of-view (FoV) of each contrast. More precisely, we created a mask of the FoV of each contrast by dilating the spinal cord segmentation in the axial plane to get the complete superior-inferior coverage. Then, we registered these FoV masks to the common T2w space (see Figure 6.1, step 4), such that the contrasts with overlapping FoVs are weighted more in the unified soft GT. Then, the averaged soft segmentation was brought back to each contrast's native space by applying the inverse warping field and using linear interpolation. This step was important because having the GT in the native space of each contrast eliminates biases due to various resolutions and fields-of-view during training. The original images along with the soft GT masks (both in their native space), were then used during training.

The vertebral levels were automatically labeled using SCT's `sct_label_vertebrae` command on the T1w and T2w images. Quality control was done using the `sct_qc` command and when necessary, labels were manually created at the posterior tip of each intervertebral disc. Since the contrasts T2*w, MT-on, GRE-T1w and DWI are axial acquisitions with thick slices, the manual or automatic labeling of the discs is not reliable. To generate the vertebral labels for

those contrasts, we warped the T2w intervertebral discs to each contrast's native space using the generated warping field in the previous step. Finally, we computed spinal cord CSA on the soft average segmentation GT and the binary segmentations averaged over the C2-C3 vertebral levels using SCT's `sct_process_segmentation`.

5.2.3 Training Protocol

In this section, we describe the preprocessing, data augmentation, our proposed model and the training strategy used for the contrast agnostic segmentation of the spinal cord.

Preprocessing

All data were resampled to 1mm isotropic resolution and re-oriented to right-posterior-inferior (RPI) before training. The images and the GT labels were resampled using spline interpolation and linear interpolation, respectively. The min, median, and max shapes of all the images in the training set after resampling were $45 \times 31 \times 70$, $192 \times 230 \times 106$, and $230 \times 320 \times 320$, respectively. As a consequence of having images with different orientations (3D, axial) and fields-of-view (cervical, cervico-thoracic, thoracic), we found center cropping to be extremely useful. Notably, the images were heavily cropped in the R-L and A-P directions to keep the spinal cord at focus while the S-I direction was left uncropped. The final patch size for center-cropping was set to $64 \times 192 \times 320$.

Data Augmentation

Given the heterogeneity across contrasts, heavy data augmentation was crucial for the performance of the model. All data augmentation transforms are random, applied with a predefined probability and called in the following order: affine transformation with spline and linear interpolation for images and labels respectively ($p = 0.9$) and the rotation, scaling and translation parameters ranging between $[-20, 20]$, $[-0.2, 0.2]$, and $[-0.1, 0.1]$, respectively, elastic deformation ($p = 0.5$) by sampling a grid of random offsets within $[25, 35]$ and Gaussian smoothing the grid with the standard deviation (STD) between $[3.5, 5.5]$, simulation of low resolution ($p = 0.25$) with a downsampling and upsampling factors sampled uniformly from $[0.5, 1.0]$, gamma correction ($p = 0.5$) with magnitude between $[0.5, 3.0]$, where 1.0 gives the original image and smaller/larger value makes image lighter/darker, respectively, bias field adjustment ($p = 0.3$) with the range of random coefficients between $[0.0, 0.5]$, Gaussian noise addition ($p = 0.1$) with mean 0.0 and the STD spread uniformly between $[0.0, 0.1]$, Gaussian smoothing ($p = 0.3$) with the STD of the smoothing kernel ranging from $[0.0, 2.0]$

for all axes, intensity scaling ($p = 0.15$) by multiplying in the range $[-0.25, 1.0]$, random mirroring ($p = 0.3$) (across all axes). Lastly, all images were normalized (independently) using z-score normalization by subtracting the mean intensity and dividing by standard deviation intensity. These augmentation transforms are readily implemented in MONAI [207].

Model Architecture

Given the popularity of the nnUNet [173], we used the same architectural template found in nnUNet's `3d_fullres` model[4]. Each layer in the encoder and decoder contains two blocks, each consisting of a convolutional layer, instance normalization [208] and leakyReLU nonlinearity [209]. Strided convolutions are used for downsampling while transposed convolutions are used for upsampling. Additionally, the network is trained with deep supervision [210], where auxiliary losses from the feature maps at each upsampling resolution are added to the final loss. This allows for the gradients to be injected deeper into the network, thus facilitating the training of all layers. The encoder made up of 5 layers, starting with 32 feature maps at the initial layer and ending with 320 feature maps at the bottleneck (i.e. $32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 320$).

Unlike nnUNet, which uses softmax activation on the logits, we followed the SoftSeg approach [15] and used normalized ReLU (NormReLU) as the final activation function. This choice is made from the observation that activation functions like sigmoid and softmax have a polarizing effect that undesirably shorten the range of soft values that carry valuable partial volume information at the boundaries. NormReLU simply normalizes the output of the ReLU activation using the maximum value, which is given by:

$$\text{NormReLU}(x) = \begin{cases} \frac{\text{ReLU}(x)}{\max \text{ReLU}(x)} & \text{if } \max \text{ReLU}(x) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

This offers the advantage of preserving the useful properties of the ReLU activation function while ensuring that the predictions are normalized within the range of 0 and 1.

Loss Function

An issue with the commonly used DiceLoss [134] is that it yields segmentation masks with sharp edges [211]. It has also been shown the optimizing with soft Dice leads to volumetric biases (due to under-/over-segmentation) with high inherent uncertainty [212]. More

[4]https://github.com/MIC-DKFZ/dynamic-network-architectures/unet.py

specifically, for our contrast agnostic segmentation problem, DiceLoss does not drive the model towards optimizing for accurate segmentations at the boundary between the spinal cord and the cerebrospinal fluid, once it obtains a *good enough* segmentation of the spinal cord. This was also observed by [213] in the context of left atrium segmentation in cardiac images. Therefore, the inability of DiceLoss to adapt its behaviour upon reaching closer to convergence (i.e., failing to distinguish between the spinal cord/CSF boundary), is the primary factor contributing towards the differences in CSA across contrasts. Although not as severe as just the DiceLoss, the same applies to other loss functions such as the Focal [133], Tversky [214], and Dice-Cross Entropy (DiceCE) losses.

As suggested in [15], we considered the segmentation task as a pixel-wise regression problem (instead of classification) and trained our model with adaptive wing loss [215]. An immediate advantage is that a regression objective produces outputs with proper calibration while allowing soft outputs lying in $[0-1]$. Originally proposed in the context of alignment of facial landmarks via heatmap regression, there are two reasons why adaptive wing loss is a suitable candidate for obtaining soft segmentations. *First*, in heatmap regression, the model regresses against the GT heatmap generated by plotting Gaussian distributions centered at each facial landmark. The mode of the Gaussian (i.e. the landmark) and the pixels in its immediate neighbourhood are considered as foreground, while the rest is background. In our case, this presents a similar class imbalance problem where the pixels at the spinal cord/CSF boundary are outnumbered by pixels at the core of the spinal cord. *Second*, loss functions assigning equal weights to all pixels during training (such as DiceLoss) do not result in accurate predictions at the boundaries. Moreover, pertaining to medical image segmentation, adaptive logarithmic losses have been shown to converge faster and mitigate class imbalance [15, 216].

The loss function is defined as follows:

$$\text{AWing}(y, \hat{y}) = \begin{cases} \omega \ln\left(1 + \left|\frac{(y-\hat{y})}{\epsilon}\right|^{(\alpha-y)}\right) & \text{if } |(y-\hat{y})| < \theta \\ A\,|(y-\hat{y})| - C & \text{otherwise,} \end{cases} \tag{5.2}$$

where y and $\hat{y}$ correspond to GT and the predicted labels, and ω, ϵ, θ, α are the hyperparameters. Briefly, the piece-wise loss function has non-linear and linear parts. The former ensures that error between the prediction and GT smaller than θ have a larger influence in the loss function (via larger gradients during backpropagation), while the latter makes the loss function behave like the mean-squared error loss with equal weights to all voxels. The definitions of the adaptive factor A, the constant term C and the hyperparameters can be found in Section 4.2 of [215]. The following hyperparameter values $\omega = 8.0$, $\epsilon = 1.0$, $\theta = 0.5$, $\alpha = 2.1$ were set for training. We also experimented with $\omega = 12.0$, $\epsilon = 0.5$ (i.e. larger ω and

smaller ϵ) as suggested in their original work, but did not observe substantial improvement in performance.

Hyperparameters & Training Details

We used the Adam optimizer [217] with a learning rate of 0.001 and a cosine annealing scheduler. The model was trained for a maximum of 200 epochs, and the batch size was set to 2. The patch size for training and sliding window inference was set to $64 \times 192 \times 320$, same as the center cropping size. All the models were trained using the MONAI [207] and PyTorch Lightning[5] frameworks on a single 48 GB NVIDIA A6000 GPU.

5.2.4 Evaluation Protocol

In this section, we describe the evaluation protocol to assess the model's performance.

Evaluation Metrics

To quantitatively evaluate the segmentation accuracy, we computed the Dice coefficient (on the binarized predictions thresholded at 0.5), average surface distance (ASD), and relative volume error (RVE) for each contrast across all test participants. To assess the variability of CSA across contrasts, we computed CSA averaged over C2-C3 vertebral levels of the cervical spinal cord on all the test set predictions for each evaluated model. The following metrics were used for quantitative evaluation:

1. *STD CSA*: The standard deviation (STD) of CSA across contrasts for each participant to assess CSA variability,
2. *Absolute CSA Error*: The absolute error between the CSA of GT segmentation and the prediction for each participant.

Mathematically, the absolute CSA error ϵ is given by:

$$\epsilon = \mid y_{\text{CSA}} - \hat{y}_{\text{CSA}} \mid, \tag{5.3}$$

where y_{CSA} corresponds to the CSA of the GT segmentation mask and $\hat{y}_{\text{CSA}}$ to the CSA of the prediction averaged at C2-C3 vertebral levels. The metrics are computed in mm^2 and the lower the STD and absolute CSA errors the better is the model, with the underlying assumption being that one participant should have the same spinal cord CSA across contrasts.

[5] https://lightning.ai

Baselines

We evaluated our model against 3 baselines, each with a different training strategy as described below.

Soft vs. Hard ground truth: To assess the impact of the type of GT mask used for training and its consequential effect on CSA variability, we trained two models: one with the soft GT generated using the procedure described in Section 5.2.2, and the second model with the contrast-specific hard GT generated using `sct_deepseg_sc` and manually corrected as required.

Single model for all contrasts vs. Contrast-specific models: Using the soft GT masks, we trained one model per contrast (i.e. 6 models) and compared it against a single model trained on all 6 contrasts. This experiment is useful in understanding whether a single model, exposed to all contrasts during training, is capable of mitigating the CSA bias.

DiceCE loss vs. Adaptive wing loss: As mentioned in Section 5.2.3, optimizing only for the Dice coefficient is insufficient for accurate segmentations at the spinal cord / cerebrospinal fluid boundary. For empirically validating this hypothesis, we treated the loss function as a hyperparameter and trained two models with: (i) Dice-Cross entropy loss, and (ii) adaptive wing loss (the proposed model) while using both the soft and hard GT segmentations.

Among all baselines, the performance of the models were evaluated in terms of the STD CSA and the absolute CSA errors across all contrasts for each participant in the test set. Except for the ablation comparing the loss functions, all other models were trained with the adaptive wing loss.

Comparison with the state of the art (SOTA)

We also compared our model's performance with a few SOTA methods, adapting it for spinal cord segmentation wherever necessary.

PropSeg: `PropSeg` [174] is based on the iterative propagation of deformable models for spinal cord segmentation. The algorithm consists of three steps: (i) an initialization step for detecting and orientating the position of the spinal cord using a circular Hough transform, (ii) a propagation step that initializes a deformable model for its propagation along the spinal cord, and (iii), a refinement step for robust and accurate segmentation of the spinal cord.

DeepSeg: `DeepSeg` [17], implemented in SCT as `sct_deepseg_sc`, features a two-stage process: (i), the spinal cord centerline is detected using a 2D CNN with dilated convolutions, and (ii), the cord is segmented along the centerline using a 2D or 3D CNN with standard

convolutions. This model was trained on 'real world' retrospective data from 30 sites including both healthy participants and pathological patients. Images were acquired from various vendors (Siemens, GE, Philips) and included 4 contrasts (T1w, T2w, T2*w and DWI) with a variety of image resolutions and fields-of-view (axial and sagittal). Because of its robustness to multi-site data, `DeepSeg` is an appropriate benchmark method.

nnUNet: The nnUNet framework [173] is the SOTA in various segmentation tasks across several challenges. We used the latest version of nnUNet (i.e. `nnUNetv2`) and train both 2D and 3D variants with the default, self-configured parameters on a single fold for 1000 epochs using all contrasts together and soft GT segmentations binarized using a threshold of 0.5. This was done because nnUNet does not yet support training with soft GT labels.

SoftSeg: SoftSeg [15] showed that by skipping the binarization step after data augmentation, one can obtain the soft labels "for free" and training on these soft GT results in better generalization and calibrated models. Contrary to our approach of creating soft labels by averaging the segmentations of multiple input contrasts *and* applying the data augmentation transforms directly on the soft labels, SoftSeg started with hard labels and trained on the soft labels obtained implicitly after data augmentation.

BigAug: `BigAug` [67], is a data augmentation-based domain generalization approach, that applied a series of 9 stacked augmentation transforms based on image quality, appearance and spatial configuration to model domain shifts. While [67] reported generalization across sites/scanners only within a single contrast (T2w), we adapted their method to compare generalization across different contrasts. Specifically, `BigAug` was trained on a collection of all 6 contrasts with hard GT labels using Dice loss and evaluated on the basis of STD CSA, absolute CSA error and generalization to unseen contrasts.

SynthSeg: SynthSeg [69] is the SOTA method for contrast-agnostic segmentation of brain MRI scans. Starting with the original ground-truth segmentations of T1w scans, we used KMeans clustering to generate the labels for all the regions outside the spinal cord. Each T1w scan in the training set was automatically clustered into 3-10 regions and the SynthSeg model was re-trained. The output segmentations from SynthSeg were not comparable to the rest of the methods. The model performed well on T1w contrast while failing to segment the cord properly on the remaining contrasts. Hence, we report its results in Appendix B along with our re-implementation details.

Generalization to Unseen Data

As described in Section 5.2.1, the Spine Generic Public Database [16] consists of healthy participants only. To evaluate our model's ability to generalize to *real world* clinical data, we tested our model on three datasets of patients presenting various spinal cord pathologies, contrasts and/or on fields-of-view unseen during training. All patients provided written informed consent following Institutional Review Board approval and the Declaration of Helsinki.

Traumatic Spinal Cord Injury (`sci-t2w`): This dataset consists of axial thoraco-lumbar T2w images of 80 patients with chronic traumatic spinal cord injury from the University of Colorado Anschutz Medical Campus. Acquisition was performed using MRI systems from 2 vendors (Siemens: $n = 16$, GE: $n = 63$) with 2 different field strengths (3T: $n = 17$, 1.5T: $n = 62$) and image resolutions ranging between $\{0.31 - 0.78\} \times \{0.31 - 0.78\} \times \{3 - 6\}$ mm^3. The challenge for the model is to be able to segment the spinal cord, in the presence of spinal cord compression, broken vertebrae and hyperintense lesions (likely edema).

Multiple Sclerosis (`ms-mp2rage`): This dataset consists of sagittal MP2RAGE "UNI" images ($1 \times 1 \times 1$ mm resolution) of 103 healthy controls and 180 multiple sclerosis patients with visible lesions from the University of Basel acquired on a Siemens MRI scanner. The challenge for the segmentation model here is that the MP2RAGE contrast is unseen during training, and that the hypointense lesions can lead to under-segmentation mainly due to the similar signal intensity as the surrounding cerebrospinal fluid.

Cervical Radiculopathy (`radiculopathy-epi`): This dataset consists of resting state axial gradient-echo echo-planar-imaging (GRE-EPI) images ($0.89 \times 0.89 \times 5$ mm resolution) of 24 participants with cervical radiculopathy and 28 age- and sex-matched healthy controls from Stanford University acquired on a GE MRI scanner. This dataset was acquired in the context of a resting state functional MRI experiment, and consists of 245 volumes that were motion corrected and averaged. Cervical radiculopathy is characterized by degenerative changes to the cervical spine, which can compress the spinal nerve roots and compromise the normal anatomy of the spinal cord, and the T2*w EPI images that can lead to strong image distortions and signal dropout making the segmentation difficult.

Ablations on the number of contrasts

As six contrasts could be considered more than what are typically acquired in MRI examinations, we performed two more experiments ablating the number of contrasts in the preprocessing and training stages, and evaluated its downstream effect on the reduction of the CSA variability across all 6 contrasts. Starting with $n = 2$ contrasts (T1w and T2w),

we followed the same preprocessing pipeline described in Section 5.2.2 and trained a model on the soft masks generated by averaging the T1w and T2w contrasts together. The same experiment was repeated for $n = 4$ contrasts (T1w, T2w, DWI, T2*w). The results are reported in Appendix B.

5.3 Results

In this section, we present the results from our proposed contrast-agnostic spinal cord segmentation model (Section 5.3.1) and evaluate them against the baselines (Section 5.3.2) and the existing SOTA methods (Section 5.3.3). Then, we show the generalization capabilities of our model on unseen, out-of-distribution data (Section 5.3.4). Lastly, in Section 5.3.5, we compare the CPU inference times between various methods.

In all the plots in the following sections, the proposed model is denoted by `soft_all`, meaning that the model was trained with a soft GT averaged from the individual segmentations of each of the 6 contrasts *and* adaptive wing loss was used as the loss function.

5.3.1 Contrast-agnostic spinal cord segmentation

Table 5.1 shows the quantitative results for the proposed contrast-agnostic spinal cord segmentation model `soft_all`. For each contrast, we present the mean ± standard deviation across test participants for Dice coefficients, relative volume errors (in %), and average surface distances. While the Dice coefficients are consistent across all contrasts, we note a slight under-segmentation in the case of MT-on and DWI contrasts (reflected by the negative RVE) and an over-segmentation for the T1w, T2w, T2*w and GRE-T1w contrasts.

Table 5.1 Quantitative results for spinal cord segmentation across contrasts on the test set (49 participants) for our `soft_all` model. RVE stands for Relative Volume Error and ASD stands for Average Surface Distance.

Contrasts	Dice (↑)	RVE %	ASD (↓)
	Opt. value: 1	Opt. value: 0	Opt. value: 0
T1w	0.96 ± 0.02	1.74 ± 3.38	0.08 ± 0.25
T2w	0.96 ± 0.01	1.89 ± 2.35	0.01 ± 0.07
T2*w	0.96 ± 0.01	0.56 ± 2.94	0.01 ± 0.01
MT-on	0.96 ± 0.02	-0.59 ± 2.88	0.01 ± 0.03
GRE-T1w	0.95 ± 0.02	0.99 ± 5.58	0.04 ± 0.09
DWI	0.96 ± 0.02	-1.04 ± 3.89	0.00 ± 0.00

Figure 5.2 shows the violin plot with absolute CSA error between the predictions and the GT across 6 contrasts (the lower the better). The mean CSA error is less than 2 mm^2 across all contrasts, which is encouraging given that 2 mm^2 represents only two pixels at an axial resolution of 1×1 mm.

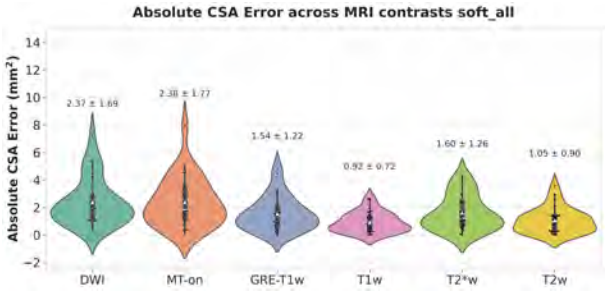

Figure 5.2 Absolute CSA error between the predictions and GT across each contrast for the proposed model. Scatter plots within each violin represent the individual CSA errors for all participants in the test set. White triangle marker shows the mean CSA error across participants.

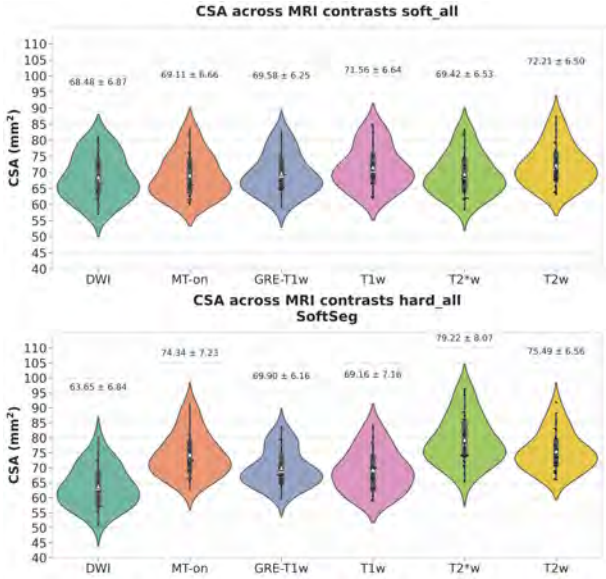

Figure 5.3 Effect of GT segmentation type (soft vs. hard) on CSA across contrasts. White triangle marker shows the mean CSA across participants.

Figure 5.3 shows the comparison of CSA across contrasts between two models trained with soft GT (top panel) and hard GT (bottom panel). Training with soft GT resulted in similar CSA across all contrasts, while the hard GT training resulted in drastically uneven distributions of CSA. Note that training with contrast-specific hard GT masks with adaptive wing loss is precisely the training strategy used in SoftSeg [15], hence we denote this model as `hard_all_SoftSeg`. We conducted a one-way paired ANOVA on CSA across contrasts for both `soft_all` and `hard_all_SoftSeg` models. The MRI contrast had a significant effect on CSA for both methods ($p < 0.05$). A follow-up posthoc analysis (two-sided Bonferroni-corrected non-parametric Wilcoxon signed-rank test) revealed that for `soft_all`, T2w / T2*w, T2*w / T1w, T1w / GRE-T1w, T2w / GRE-T1w, T2w / MT-on, T2w / DWI, T1w / MT-on, and T1w / DWI pairs of contrast showed a significant difference between CSA ($p < 0.05$). While for `hard_all_SoftSeg`, all pairs of contrasts show significantly different CSA values ($p < 0.05$) except for the T2w/MT-on pair.

Despite significant paired differences across contrasts for both `soft_all` and `hard_all_SoftSeg`

models, the variability of CSA across contrasts did indeed reduce significantly, as described in the next section.

5.3.2 Comparison with baselines

Figure 5.4 compares the performance of the `soft_all` model with the baselines in terms of the STD of CSA across contrasts. The STD is computed across all 6 contrasts for each participant and each test participant is represented by an individual point in the scatter plot. Starting with the GT violin plots on the left (gray), we observe that the root cause of CSA variability can be mitigated using soft average segmentations as the GT during training. This is also supported by the fact that the `hard_all_SoftSeg` model, trained on binary GT, results in higher STD when compared to its soft counterparts. Within the models trained using soft GT, the performance of a single model trained on all contrasts (`soft_all`) is similar to the average of 6 models trained individually on each contrast (`soft_per_contrast`). Lastly, on comparing the effect of training with DiceCE loss (`soft_all_diceCE_loss`) and adaptive wing loss (`soft_all`), we observe significantly lower CSA STD across contrasts when using the regression-based adaptive wing loss ($p < 0.001$).

While the previous figure showed the STD of CSA across contrasts for each of the test participant, Figure 5.5 compares the absolute CSA error between the prediction and the GT for our model and the baselines. The points in the scatter plots represent each test image and the mean CSA error is given on top of the violin plots. The superior performance of `soft_all` suggests that a combination of soft segmentations GT along with adaptive wing loss is crucial for mitigating CSA variability. When comparing the CSA errors between `soft_per_contrast` and `soft_all` models, we observe significantly lower CSA errors ($p < 0.001$) with the latter as also depicted in the violin plot containing a high density of scatter points between $0-2$ mm^2 range. More plots comparing the variability in absolute CSA errors across each contrast between the baselines are shown in Appendix B.

5.3.3 Comparison with the state of the art

Given the results from the comparison with baselines in the previous section, we considered `soft_all` as the best model for subsequent comparisons with the existing SOTA methods. Recall that `soft_all` denotes the model that has been trained with soft GT averaged from the segmentations of each of the 6 contrasts *and* adaptive wing loss was used as the loss function. Table 5.2 shows the quantitative results for the proposed segmentation model (`soft_all`), and the existing SOTA methods. The `soft_all` model outperforms the SOTA methods in terms on Dice coefficient and performs slightly worse in terms of RVE compared to DeepSeg2D

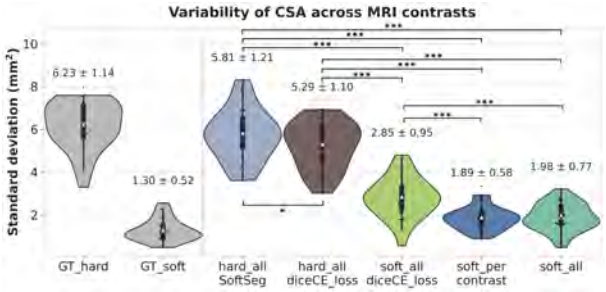

Figure 5.4 Standard deviation of CSA averaged across C2-C3 vertebral levels compared to the baselines (the lower the better). `hard_all_SoftSeg` refers to the single model trained using all contrasts with hard GT and the SoftSeg training approach [15], `hard_all_diceCE_loss` refers to the single model trained with the DiceCE loss and hard individual GT, `soft_all_diceCE_loss` refers to the single model trained with the Dice CE loss and soft GT, `soft_per_contrast` refers to the mean of 6 individual models trained on 6 contrasts with soft GT, and `soft_all` refers to the single model trained using all contrasts with soft GT. White triangle marker shows the mean. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-sided Bonferroni-corrected non-parametric Wilcoxon signed-rank test).

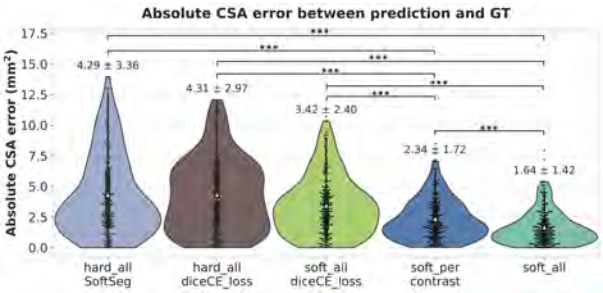

Figure 5.5 Mean absolute CSA error compared against the baselines. `hard_all_SoftSeg` refers to the single model trained using all contrasts with hard GT and the SoftSeg training approach [15], `hard_all_diceCE_loss` refers to the single model trained with the Dice CE loss and hard individual GT, `soft_all_diceCE_loss` refers to the single model trained with the DiceCE loss and soft GT, `soft_per_contrast` refers to the mean of 6 individual models trained on 6 contrasts with soft GT, and `soft_all` refers to the single model trained using all contrasts with soft GT. White triangle marker shows the mean. ** $p < 0.01$, *** $p < 0.001$ (two-sided Bonferroni-corrected non-parametric pairwise Wilcoxon signed-rank test between `soft_all` and the 4 other methods.

and average surface distance (ASD) compared to nnUNet2D, respectively. DeepSeg2D and both nnUNet models show under-segmentation (reflected by the negative RVE value) while PropSeg, DeepSeg3D, BigAug and SoftSeg models present over-segmentation. Also note the relatively low Dice STD for the `soft_all` model, suggesting higher robustness (i.e., other models fail more often in presence of difficult images).

Figure 5.6 compares the STD of CSA across contrasts for our best model and the existing methods. DeepSeg2D has the highest STD across contrasts, followed by `hard_all_SoftSeg` and `hard_all_BigAug`.

Interestingly, despite the hard requirement of having binarized GT and training with DiceCE loss, both nnUNet models achieved similar STD across contrasts with the 3D model showing lower STD across contrasts. Overall, both nnUNet 2D and 3D models showed higher STD when compared to our best model `soft_all`.

Table 5.2 Quantitative comparison of spinal cord segmentations for the state of the art methods on the test set (294 images) averaged across all contrasts. 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
PropSeg [174]	0.85 ± 0.15	7.18 ± 32.59	0.49 ± 3.92
DeepSeg3D [17]	0.85 ± 0.13	18.25 ± 49.12	0.12 ± 0.31
`hard_all_BigAug` [67]	0.92 ± 0.02	3.16 ± 6.07	0.02 ± 0.10
DeepSeg2D [17]	0.95 ± 0.03	$\mathbf{-0.24 \pm 8.64}$	0.06 ± 0.31
nnUNet3D [173]	0.95 ± 0.02	-2.11 ± 4.43	0.04 ± 0.29
nnUNet2D [173]	0.95 ± 0.02	-1.85 ± 4.46	0.02 ± 0.10
`hard_all_SoftSeg` [15]	0.96 ± 0.02	1.77 ± 5.74	$\mathbf{0.01 \pm 0.05}$
`soft_all` (ours)	$\mathbf{0.96 \pm 0.01}$	0.6 ± 3.82	0.03 ± 0.12

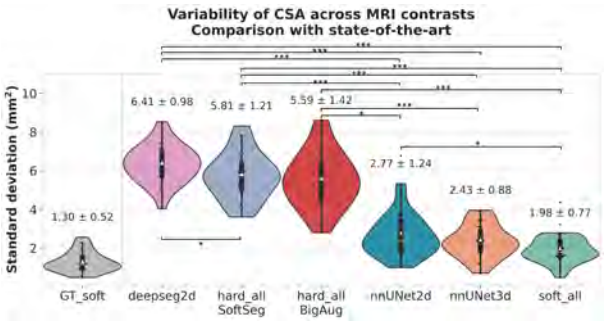

Figure 5.6 Standard deviation of CSA between C2-C3 vertebral levels for DeepSeg2D, `hard_all_SoftSeg`, `hard_all_BigAug`, nnUNet 2D/3D, and our model `soft_all`. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-sided Bonferroni-corrected non-parametric pairwise Wilcoxon signed-rank test between each pair of methods.

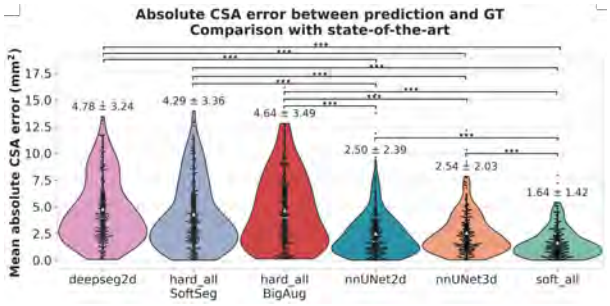

Figure 5.7 Mean absolute CSA error for DeepSeg 2D, `hard_all_SoftSeg`, `hard_all_BigAug`, nnUNet 2D/3D, and our model `soft_all`. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-sided Bonferroni-corrected non-parametric pairwise Wilcoxon signed-rank test between each pair of methods.

Figure 5.7 shows a comparison between our best model (`soft_all`) and the other methods in terms of the absolute CSA error. Similar to the trend observed in Figure 5.6, we noted that `soft_all` achieves the lowest mean absolute CSA error with 1.64 ± 1.42 mm^2 across all test images. In Figures B.5 and B.6, we show the comparison with *all* the methods including DeepSeg 3D and PropSeg to ovoid overcrowding Figures 5.6 and 5.7.

To understand the impact of data augmentation and loss functions on training with soft masks, we compared our models trained using DiceCE and adaptive wing loss with nnUNet in Figure 5.8. Keeping the loss function fixed (i.e. DiceCE), we can observe that data

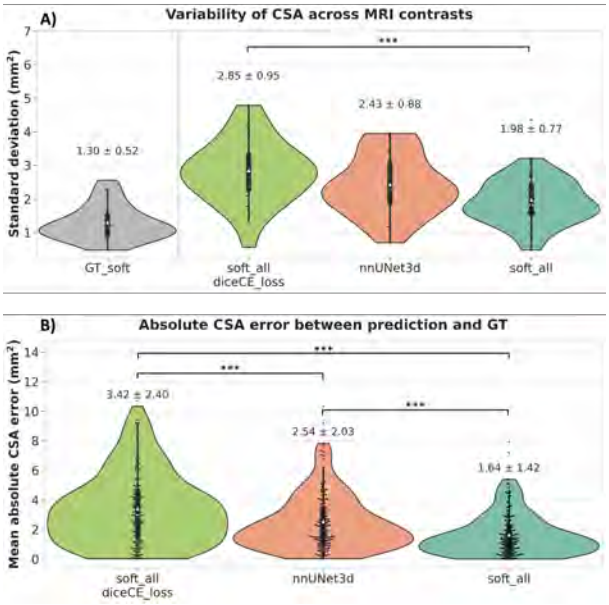

Figure 5.8 Comparison of CSA estimation between models trained on soft masks. **A)** Standard deviation of CSA between C2-C3 vertebral levels for `soft_all_diceCE_loss`, nnUNet 3D, and our model `soft_all`. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-sided Bonferroni-corrected non-parametric pairwise Wilcoxon signed-rank test between `soft_all`, nnUNet 3D and `soft_all_diceCE_loss`). **B)** Mean absolute CSA error for `soft_all_diceCE_loss`, nnUNet 3D, and our model `soft_all`. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (two-sided Bonferroni-corrected non-parametric pairwise Wilcoxon signed-rank test between `soft_all`, nnUNet 3D and `soft_all_diceCE_loss`.

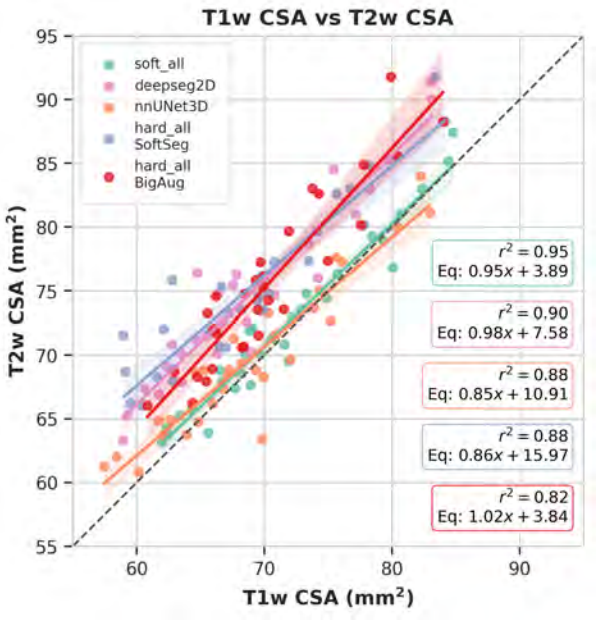

Figure 5.9 Level of agreement between T1w and T2w CSA for the best-performing SOTA methods. Each point represents one participant. The black dashed line represents perfect agreement between the CSA of T1w and T2w contrasts.

augmentation transforms in nnUNet3D play a stronger role as it achieves lower STD and absolute CSA errors compared to the `soft_all_diceCE_loss` model. However, as shown by `soft_all`, switching to the adaptive wing loss irrespective of data augmentation transforms leads to further reduction in the CSA variability across contrasts.

In Figure 5.9, we plotted the level of agreement between the CSA estimated by our model (`soft_all`) and the SOTA methods on T1w and T2w contrasts. The models trained with individual hard masks (namely, `deepseg2D`, `hard_all_SoftSeg`, and `hard_all_BigAug`)

showed large discrepancies between the estimated CSA. Interestingly, the models trained with soft masks obtained by averaging all contrasts moved closer to the diagonal line representing perfect alignment between T1w/T2w CSA. Within `soft_all` and `nnUNet3D`, our model achieves better alignment between the two contrasts, thus confirming that it reduces discrepancies between these two popular contrasts in spinal cord imaging. The correlation plots for the remaining pairs of contrasts for `soft_all` are shown in Figure B.1 in Appendix B.

5.3.4 Generalization to unseen data

Figure 5.10 shows the predictions for 8 representative patients with spinal cord injury from the `sci-t2w` dataset. Despite the presence of spinal cord lesions, we notice that `soft_all` and nnUNet models were able to correctly segment the spinal cord, while DeepSeg2D T2w and `soft_per_contrast` T2w models under-segmented the spinal cord (except one over-segmentation pointed by the red arrow). The models `hard_all_SoftSeg`, `hard_all_BigAug`, `hard_all_diceCE` and `soft_all_diceCE` showed under-segmentation of the hyperintense lesions in the spinal cord as indicated by the red arrows.

Figure 5.11 shows the predictions for 6 representative patients with multiple sclerosis and 2 healthy participants from the `ms-mp2rage` dataset. Despite the presence of spinal cord MS lesions, we observe that the `soft_all` model was able to correctly segment the spinal cord, while DeepSeg2D T1w model under-segmented the spinal cord typically at the lesion location. The `soft_per_contrast` T1w and nnUNet models were unable to properly segment the spinal cord in the presence of hypointense lesions. `hard_all_BigAug` showed some under-segmentation of the location of the hypointense lesions. `hard_all_SoftSeg`, `hard_all_diceCE` and `soft_all_diceCE` showed similar performance and were able to properly segment the spinal cord.

Figure 5.12 shows the predictions for 4 representative patients with cervical radiculopathy and 4 healthy controls from the `radiculopathy-epi` dataset. The `soft_all` model was able to correctly segment the spinal cord even with the poor image quality of the gradient echo EPI. In contrast, the `soft_per_contrast` T2*w was unable to segment the spinal cord in almost all cases. The DeepSeg2D T2*w and nnUNet performed slightly worst than `soft_all`, notably in slices affected by signal drop out (e.g., HC013), and nnUNet, `hard_all_diceCE`, `hard_all_SoftSeg` had a tendency to leak into the cerebrospinal fluid (red arrows). The `hard_all_BigAug` model had trouble with getting the shape of the spinal cord, mainly in the presence of signal drop-out, shown by the red arrow.

Table 5.3 presents the quantitative metrics for the models' performance on all three *external*

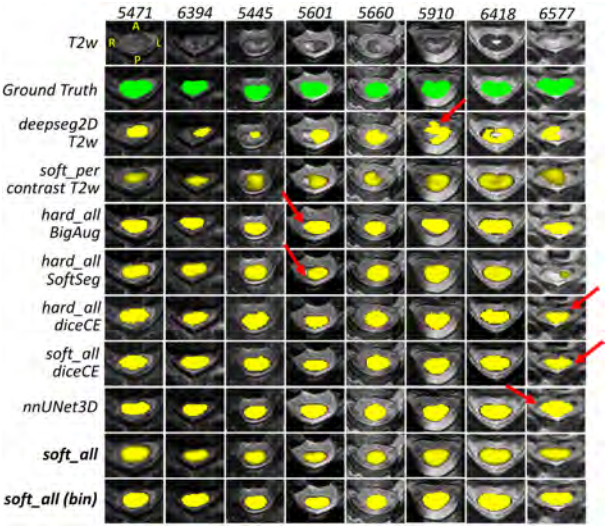

Figure 5.10 T2w axial slices with the overlaid GT (green) and model predictions (yellow) in 8 patients with traumatic spinal cord injury (`sci-t2w` dataset). Red arrow depict segmentation errors. Soft segmentations are clipped at 0.5. `soft_all(bin)` represents the `soft_all` binarized at 0.5 for better visual comparison with the GT and hard segmentation methods.

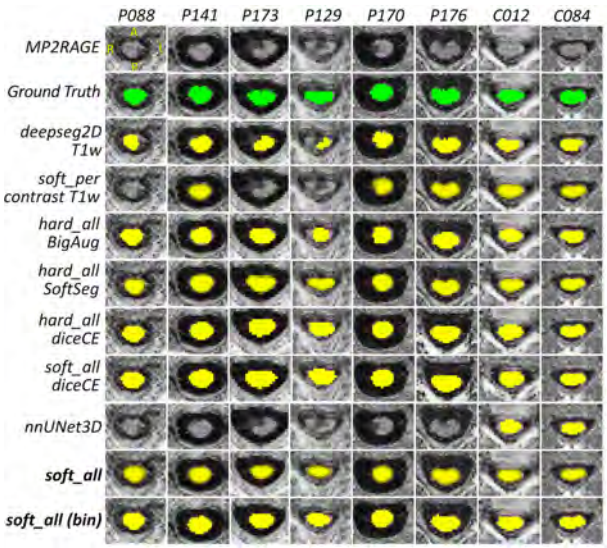

Figure 5.11 MP2RAGE axial slices with the overlaid GT (green) and model predictions (yellow) in 6 patients (P) with multiple sclerosis lesions and 2 healthy controls (C) (`ms-mp2rage` dataset). Soft segmentations are clipped at 0.5. `soft_all(bin)` represents the `soft_all` binarized at 0.5.

datasets. We do not report the Dice coefficients for DeepSeg2D since the GT masks were generated using this model with manual corrections on the slices presenting over/under-segmentations, hence biasing the scores. We observed that the nnUNet3D performed well on the T2w and GRE-EPI contrasts while performing poorly on the MP2RAGE "UNI" contrast (Dice = 0.24 ± 0.25) due to large under-segmentation overall (RVE = −82.51 ± 19.27%) while both models using DiceCE (`hard_all_DiceCE` and `soft_all_DiceCE`) performed well on MP2RAGE "UNI" contrast, as we can also observe in Figure 5.11. Both BigAug [67] and SoftSeg [15] models outperformed nnUNet3D only for the MP2RAGE contrast. The `soft_per_contrast` model performed poorly on GRE-EPI (Dice = 0.29 ± 0.18) but perfomed well on the MP2RAGE "UNI" and T2w data. In terms of RVE, we see that nnUNet 3D consistently shows under-segmentation on all datasets. The `soft_all` model outperforms all the tested methods for `radiculopathy-epi` datasets in terms of Dice coefficients and performs similar to nnUNet on the `sci-t2w` dataset, and similar to SoftSeg on the `ms-mp2rage` dataset.

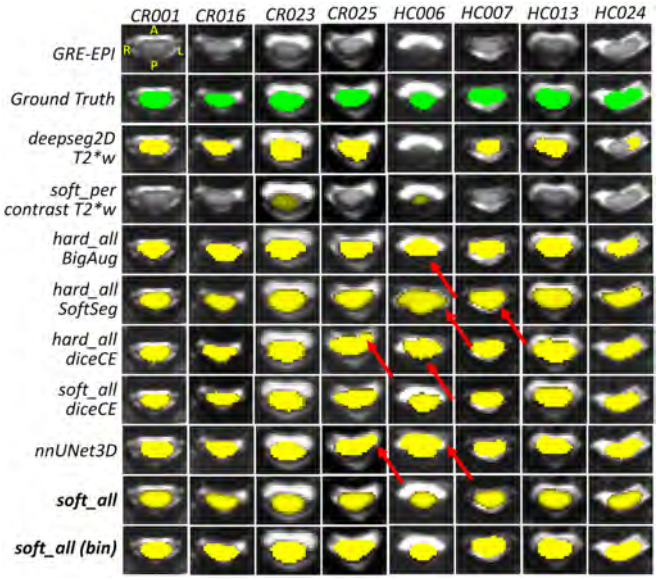

Figure 5.12 GRE-EPI axial slices with the overlaid GT (green) and model predictions (yellow) of spinal cord segmentation of 4 patients with cervical radiculopathy (CR) and 4 healthy controls (HC). Soft segmentations are clipped at 0.5. `soft_all(bin)` represents the `soft_all` binarized at 0.5. Red arrows indicate examples of segmentation errors.

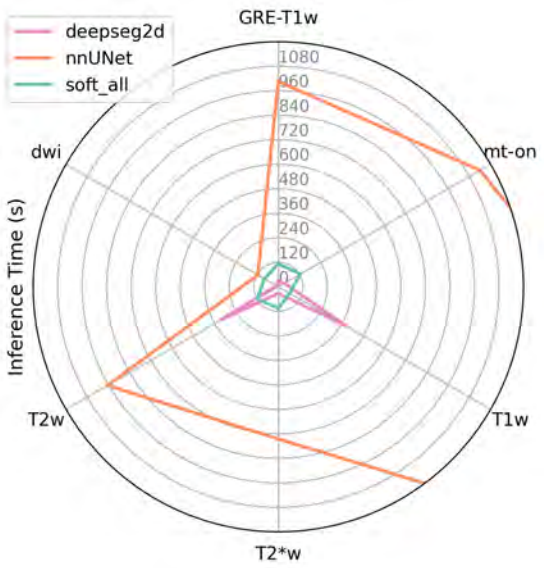

Figure 5.13 Inference times (in seconds) averaged across test participants for DeepSeg2D, nnUNet3D, and `soft_all` for all contrasts.

5.3.5 Inference times

Figure 5.13 compares the average inference time across participants for the DeepSeg2D, nnUNet3D and `soft_all` methods. The inference is done on a CPU (Intel Xeon E7-4850 @ 2.10GHz) and the time is shown in seconds. `soft_all` takes up to 2 minutes per prediction on average irrespective of the contrast, whereas nnUNet 3D's inference times are highly variable and longer. For example, nnUNet3D takes about 3000 seconds (50 mins) for segmenting a T1w image (not shown in the plot for clarity), while `soft_all` requires only about 2 minutes. Unsurprisingly, the inference time largely depends on the size of the input volume. In addition to obtaining good segmentations, the average inference time per image is an important factor for consideration before deploying the model in real-world clinical settings. Models such as nnUNet3D requiring long inference times are impractical when used on large cohorts.

Table 5.3 Comparison of quantitative metrics between SOTA methods for spinal cord segmentation on unseen datasets. n refers to the number of participants.

Methods	Dice (↑)	RVE %	ASD (↓)
	Opt. value: 1	Opt. value: 0	Opt. value: 0
	Dataset `sci-t2w` ($n = 80$)		
`hard_all_DiceCE`	0.86 ± 0.06	-8.83 ± 11.95	0.03 ± 0.12
`hard_all_BigAug`	0.83 ± 0.08	$\mathbf{-5.67 \pm 14.84}$	0.03 ± 0.16
`hard_all_SoftSeg`	0.85 ± 0.09	-14.44 ± 14.25	0.17 ± 0.98
`soft_per_contrast`	0.84 ± 0.09	-16.45 ± 11.19	0.04 ± 0.19
`nnUNet3D`	$\mathbf{0.87 \pm 0.05}$	-15.20 ± 8.59	$\mathbf{0.01 \pm 0.03}$
`soft_all_DiceCE`	$\mathbf{0.87 \pm 0.05}$	-8.30 ± 10.72	0.02 ± 0.10
`soft_all`	0.86 ± 0.07	-16.14 ± 9.80	0.01 ± 0.06
	Dataset `ms-mp2rage` ($n = 283$)		
`hard_all_DiceCE`	0.92 ± 0.03	8.51 ± 6.15	0.01 ± 0.03
`hard_all_BigAug`	0.89 ± 0.05	-2.38 ± 9.69	0.20 ± 0.75
`hard_all_SoftSeg`	$\mathbf{0.93 \pm 0.02}$	7.57 ± 5.61	0.01 ± 0.08
`soft_per_contrast`	0.83 ± 0.14	-14.38 ± 22.44	0.15 ± 0.44
`nnUNet3D`	0.24 ± 0.25	-82.51 ± 19.27	5.09 ± 18.15
`soft_all_DiceCE`	0.90 ± 0.04	17.56 ± 7.54	0.01 ± 0.04
`soft_all`	0.93 ± 0.03	$\mathbf{6.88 \pm 6.09}$	$\mathbf{0.01 \pm 0.03}$
	Dataset `radiculopathy-epi` ($n = 52$)		
`hard_all_DiceCE`	0.87 ± 0.04	20.98 ± 11.56	0.01 ± 0.02
`hard_all_BigAug`	0.87 ± 0.03	-2.52 ± 10.4	0.01 ± 0.01
`hard_all_SoftSeg`	0.88 ± 0.04	17.14 ± 10.85	0.02 ± 0.02
`soft_per_contrast`	0.29 ± 0.18	-80.64 ± 12.82	0.58 ± 1.06
`nnUNet3D`	0.90 ± 0.04	-4.66 ± 8.79	0.01 ± 0.02
`soft_all_DiceCE`	0.88 ± 0.04	10.53 ± 14.36	0.01 ± 0.01
`soft_all`	$\mathbf{0.90 \pm 0.03}$	$\mathbf{2.83 \pm 11.22}$	$\mathbf{0.01 \pm 0.01}$

5.4 Discussion and Conclusion

We presented an automatic method for the contrast-agnostic soft segmentation of the spinal cord. Starting with the creation of GT masks, we proposed a new framework for generating a unique (soft) GT that represents the segmentations across various MR contrasts. Using these soft GT masks along with the corresponding images of all contrasts as inputs, we trained a U-Net model with aggressive data augmentation, regression-based adaptive wing loss, and NormReLU as the final activation function for spinal cord segmentation. We evaluated our model against 3 categories of baseline models to assess the impact of soft vs. hard GT masks, training a single model with all contrasts vs. training one model per contrast, and the impact of the loss functions on the subsequent morphometric measures (i.e., CSA). We compared our

model's performance with the SOTA methods for spinal cord segmentation, namely, PropSeg, DeepSeg, SoftSeg, BigAug and nnUNet. To demonstrate our model's domain generalization capabilities on unseen contrasts and images with pathology (lesions), we presented a qualitative comparison against spinal cord segmentations of the SOTA methods. Lastly, we provided a graphical illustration of the average inference times of the existing methods and highlighted that our model takes only a fraction of the time per image irrespective of the contrast, while obtaining better segmentations that reduce the bias/variability in morphometric measures.

5.4.1 Preprocessing for soft GT

Our preprocessing framework to generates a unique soft segmentation from individual hard segmentations. The procedure required aligning all the images to the T2w image space, including resampling and reorientation. The soft segmentation was then brought back in each contrast's native space. Then, in order to have a fixed patch size for training with all contrasts, we performed an up-sampling to 1 mm isotropic on all images and labels as a preprocessing step during data augmentation. One could question the reason for this up-down-up sampling in the preprocessing workflow, a few points are discussed below:

- **Comparison with baselines**: In order to provide a fair comparison with the baselines, we ensured that all methods used the ground truth defined in the native space of each contrast.
- **Computation of CSA and disc labeling**: Continuing the previous point and considering the evaluation of CSA variability across contrasts, it was important to compute the CSA in the native space, notably because of the highly variable spatial resolution across contrasts, which limited our ability to ensure that vertebral coverage would be the same when computed in the native image vs. in the reference image (T2w 0.8mm isotropic).
- **Patch size**: To ensure a uniform patch size during training for all the 6 contrasts, resampling the images to a unique, common, resolution was necessary. After experimenting with various spatial resolutions (trade-off between computational resource and precision), we chose 1 mm as the target resolution.
- **Introduce realistic variability**: One of the advantages of training the model in the native space is that, for a given subject, the spinal cords are not perfectly aligned across contrasts.

5.4.2 Variability of CSA across contrasts

When comparing the performance on all 6 MR contrasts (Figure 5.2), DWI and MT-on contrasts yielded the highest absolute CSA error with 2.37 ± 1.09 mm^2 and 2.38 ± 1.77 mm^2, respectively. This can be attributed to the fact that DWI and MT-on contrasts have less well defined boundaries between the spinal cord and the cerebrospinal fluid due to the coarser resolution of the images ($0.9 \times 0.9 \times 5$ mm). Furthermore, the presence of susceptibility related artifacts on DWI and MT-on (MT-on is based on GRE readout and hence suffers from signal dropout), and/or the ghosting and motion artifacts that particularly affect MT-on data [42] could explain the larger errors.

The agreement between T1w and T2w CSA (Figure 5.9) leads to a linear equation given by $0.95x + 3.89$ with an $r^2 = 0.95$, which is very close to the (dashed) identity line. In Cohen-Adad *et al.* [42], where `sct_deeepseg_sc` was used for generating spinal cord segmentations, the authors reported poor agreement between CSA computed on T1w and T2w images, regardless of the MR vendor.

Furthermore, training an individual model for each contrast (`soft_per_contrast`) yielded similar performance (albeit with a slightly higher error for each contrast) compared to training a single model for all contrasts (`soft_all`), as seen in Figure 5.4. Training an individual model for each contrast does not help in mitigating the CSA variability across contrasts as each model is optimizing for spinal cord segmentations asynchronously, thus leading to different CSA for a given participant despite using soft GT segmentations. On the other hand, a single model trained on all contrasts together is exposed to the wide heterogeneity in the images, thus leading to better estimation of the CSA and better generalization (as shown by our results on `ms-mp2rage` and `radiculopathy-epi` datasets in Figures 5.11 and 5.12). Moreover, from a deployment standpoint, packaging and distributing a single model is more convenient and intuitive for researchers to use.

Overall, the proposed contrast-agnostic `soft_all` model outperforms the baselines and existing state-of-the-art spinal cord segmentation methods while minimizing contrast-specific CSA biases.

5.4.3 Effects of ground truth masks and loss functions

Results from Figures 5.3, 5.4, and 5.5 demonstrated that the types of GT masks and loss functions used play a crucial role in the (downstream) computation of CSA. Using the unique soft GT generated by averaging the segmentations across 6 contrasts compared to traditional hard GT leads to lower CSA variability on the predictions. For more details, see Figure 5.3

(qualitative assessment) and Figure 5.4 (statistical assessment). Notably, the bias inherent to the individual GT for each contrast is propagated *less* when using the unique soft GT averaged from different contrasts.

For `soft_all`, the CSA did not differ significantly between T1w and T2w contrasts, while it was significantly different between the T1w / T2w and the other contrasts (T2*w, GRE-T1w, MT-on, DWI). Interestingly, T1w and T2w images have similar isotropic resolutions (respectively $1 \times 1 \times 1$ mm and $0.8 \times 0.8 \times 0.8$ mm), whereas the other contrasts feature highly anisotropic axial acquisitions (> 3 mm slice thickness). It is therefore possible that excessive partial volume effect along the superior-inferior axis created a bias in the CSA estimation. Another possible explanation of the discrepancy between isotropic and anisotropic scans, is the uncertainty in the estimation of the C2-C3 vertebral labels across contrasts. Since it was not possible to directly label the vertebral levels on the anisotropic scans (because discs are poorly visible), we used the disc labels created from the T2w images and applied the warping field to the labels to the target contrast. This likely resulted in slightly higher CSA STD across contrasts.

When the Dice or DiceCE loss functions [134, 173] were used in combination with hard GT masks, our results suggested that using Dice metric in the training objective is not sufficient for achieving accurate segmentations at the spinal cord / cerebrospinal fluid boundary. In fact, the model trained with Dice loss (`hard_all_BigAug`) showed subtle under-segmentations upon quality control as supported by larger absolute CSA errors in Figure 5.5. Instead, using adaptive wing loss that switches to the logarithmic (non-linear) part of the loss when the error between the prediction and the GT are small, helps the model in refining the segmentations at the boundaries of the spinal cord, thus mitigating the CSA bias across contrasts. Similar observations have been reported about the effectiveness of regression-based [15] and logarithmic [216] loss functions.

The benefits of using soft GT and adaptive wing loss in our model `soft_all` can be seen in Figures 5.6 and 5.7. PropSeg, DeepSeg, SoftSeg and BigAug methods, which used hard GT that are inherently biased due to the procedure of their GT generation, resulted in higher STD across contrasts per participant. As nnUNet does not support soft training yet, using soft segmentations averaged from all contrasts (but binarized at 0.5 threshold), still resulted in slightly higher CSA variability. Furthermore, DeepSeg used Dice loss and nnUNet used DiceCE loss during training, thus explaining the larger errors. Thus, the subtle yet important difference of training on implicitly obtained soft masks via data augmentation vs. applying the augmentation transforms directly on the soft masks generated from multiple contrasts has a significant downstream impact on the reduction of CSA variability across contrasts.

5.4.4 Generalization to unseen data

The proposed `soft_all` model generalizes well to the unseen MP2RAGE "UNI" (`ms-mp2rage`) and resting state GRE-EPI (`radiculopathy-epi`) contrasts. Despite being trained only on healthy participants, it performed well on patients with MS lesions, traumatic spinal cord injury and cervical radiculopathy. We noticed that the tested models performed similarly on a contrast that was included during training (T2w) as reflected by the Dice coefficients of Table 5.3 and segmentations in Figure 5.10, even in the presence of traumatic spinal cord injury. Surprisingly, there is a marked difference between the qualitative segmentations of (`soft_all`) and nnUNet3D (Figures 5.11 and 5.12) on the `ms-mp2rage` and `radiculopathy-epi` datasets, respectively. With the former dataset, nnUNet3D performs poorly by completely missing the spinal cord in the presence of multiple sclerosis lesions, whereas for the latter, we observed cases with over-segmentation of the cord leaking into the cerebrospinal fluid. For both these contrasts, `soft_all` obtains accurate segmentations of the spinal cord under the presence of lesions and along the spinal cord boundaries.

The difference in the segmentations between our model and nnUNet is likely due to our improved training strategy involving cropping along the center of the spinal cord, regression-based adaptive wing loss, and most importantly, training *directly* on the soft GT masks (unlike in nnUNet where the soft GT were binarized). The localization of the spinal cord, mainly through cropping, has been a recurrent prerequisite step in the literature [17,218,219], suggesting its necessity for obtaining robust segmentations.

Furthermore, our model `soft_all` is able to better delineate the shape of the spinal cord even in the presence of lesions compared to models trained specifically on one contrast (`soft_per_contrast T2w` and DeepSeg2D). This enhanced performance can likely be attributed to the model's comprehensive exposure to diverse contrasts of the spinal cord, cerebrospinal fluid, gray matter and white matter, that are featured in the Spine Generic [16] dataset.

5.4.5 Limitations & future work

Application to thoracic and lumbar levels

The proposed model is trained on a dataset of healthy participants containing cervical and upper-thoracic spinal cords only. Future research will add images with lumbar cords to further improve the generalizability of the model towards different fields-of-view.

Center cropping

The center cropping step in the online preprocessing of the images during training and inference assumes that spinal cord is at the center of the image. While this is a reasonable assumption, there might be some edge cases containing lumbar spines or participants with scoliosis on which the automatic predictions might fail. In order to mitigate this issue, the SCT function that runs inference with the proposed model provides a flag to change the default cropping (allowing researchers to adjust the cropping sizes based on their images).

Binarization threshold

At prediction time, the soft output was binarized for comparison with other methods. That binarization was done using a 0.5 threshold. As discussed in [15], that threshold could potentially be optimized to further reduce the variability of CSA across contrasts. However, this would imply arbitrarily categorizing images into a given contrast, which defeats the purpose of the current contrast-agnostic method, wherein the model can be used as is regardless of the acquisition parameters. Moreover, with MRI acquisitions, it is difficult to define the contrasts accurately as, for instance, some combinations of parameters could lead to more/less T2w or more/less magnetization transfer saturation depending on the offset of the MT pulse and/or the presence of saturation bands.

Validation in pathologies

As mentioned in the introduction, one of the advantages of soft segmentation is the ability to encode volumetric measures with finer precision compared to binary segmentation. Considering that changes in the spinal cord happen at a very slow rate at early stages of MS, we expect that soft segmentation of the spinal cord will help detect subtle atrophies. We expect that the soft segmentation of the spinal cord will increase the precision of spinal cord CSA measurements, and thus lead to lower arm-size in trials [99, 102] . This will especially be of interest in the context of cross-sectional studies where protocols can vary. In patients with degenerative cervical myelopathy, which is characterized by a progressive compression of the spinal cord, being able to precisely segment the spinal cord could also lead to better prognostication and therapeutic strategies [187].

Continual model refinement through active learning

The improvement in the mitigation of biases in morphometric measures between MRI contrasts holds exciting prospects for future work. Thanks to the remarkable generalization of

the proposed model to unseen contrasts, datasets containing other contrasts (e.g., phase-sensitive inversion recovery, short tau inversion recovery, susceptibility-weighted imaging, MP2RAGE) can be added to the existing cohort to further improve the generalizability of the model. Moreover, enriching the model with images from patients with spinal pathologies will likely improve zero-shot generalization on participants with lesions and/or spinal cord compressions. For these advancements, human-in-the-loop active learning [57] involving an initial batch of segmentations from our model followed by manual corrections of under-/over-segmentations and then re-training the model on the larger datasets until fully automatic predictions are obtained is an attractive strategy. This approach for gradually aggregating large-scale datasets while improving the model simultaneously is similar to the recently proposed Segment Anything [220] model, thus paving way towards a foundational model for contrast-agnostic spinal cord segmentation.

5.4.6 Comparison with other model architectures

We extended this work by evaluating the contrast-agnostic segmentation capabilities of different classes of DL architectures, namely, ConvNeXt, vision transformers (ViTs) and hierarchical ViTs [221]. Using the same open-access spine-generic dataset, keeping the preprocessing and training protocols fixed, we performed a comparative study of seven different DL models. Our results showed that CNNs produced robust SC segmentations across contrasts, followed by ConvNeXt, and hierarchical ViTs. This suggests that: (i) inductive biases such as learning hierarchical feature reprensentations via pooling (common in CNNs) are crucial for good performance for spinal cord segmentation, and (ii) hierarchical ViTs that incorporate several CNN-based priors can perform similarly to pure CNN-based models. We refer the reader to Appendix C for more details.

CHAPTER 6 ARTICLE 3: MONITORING MORPHOMETRIC DRIFT IN LIFELONG LEARNING SEGMENTATION OF THE SPINAL CORD

Authors Enamundram Naga Karthik[1,2], Sandrine Bédard[1], Jan Valošek[1,2,3,4], Christoph S. Aigner[5,6], Elise Bannier[7], Josef Bednařík[8,9], Virginie Callot[10,11], Anna Combes[12,13], Armin Curt[14], Gergely David[14,15], Falk Eippert[16], Lynn Farner[14], Michael G Fehlings[17,18], Patrick Freund[14,19,20], Tobias Granberg[21,22], Cristina Granziera[23], RHSCIR Network Imaging Group[24], Ulrike Horn[16], Tomáš Horák[8,9], Suzanne Humphreys[24], Markus Hupp[14], Anne Kerbrat[25,26], Nawal Kinany[27,28], Shannon Kolind[29], Petr Kudlička[9,30], Anna Lebret[14], Lisa Eunyoung Lee[31], Caterina Mainero[32], Allan R. Martin[33], Megan McGrath[13], Govind Nair[34], Kristin P. O'Grady[13], Jiwon Oh[35], Russell Ouellette[21,22], Nikolai Pfender[14], Dario Pfyffer[36,14], Pierre-François Pradat[37], Alexandre Prat[38], Emanuele Pravatà[39,40], Daniel S. Reich[41], Ilaria Ricchi[27,28], Naama Rotem-Kohavi[24], Simon Schading-Sassenhausen[14], Maryam Seif[14,19], Andrew Smith[42], Seth A Smith[13], Grace Sweeney[13], Roger Tam[43], Anthony Traboulsee[44], Constantina Andrada Treaba[32], Charidimos Tsagkas[41,23], Zachary Vavasour[43], Dimitri Van De Ville[27,28], Kenneth Arnold Weber II[45], Sarath Chandar[2,46], Julien Cohen-Adad[1,2,47,48]

Affiliations

1. NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montréal, Montréal, Québec, Canada
2. Mila - Québec Artificial Intelligence Institute, Montréal, Québec, Canada
3. Department of Neurosurgery, Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc, Czechia
4. Department of Neurology, Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc, Czechia

   ⋮

46. Canada CIFAR AI Chair
47. Functional Neuroimaging Unit, CRIUGM, University of Montreal, Montreal, Québec, Canada
48. Centre de recherche du CHU Sainte-Justine, Université de Montréal, Montréal, Québec, Canada

This article is under review at the *Imaging Neuroscience* journal, submitted on 2025-04-14.

Complete list of affiliations for all the co-authors can be found in Appendix E.

Contributions

Naga Karthik led the project, curated the datasets, ran experiments, designed the drift monitoring pipeline, and wrote the draft of the manuscript. Sandrine Bédard and Jan Valošek helped in QCing the datasets used for training, updating the normative database of morphometric measures, and were involved in discussions and writing. Rest of the authors were involved in data acquisition at their respective sites and shared the data. Julien Cohen-Adad provided supervision throughout the project. All co-authors reviewed the manuscript.

A preliminary proof-of-concept of the lifelong learning approach presented in this work has been published as a workshop paper [222] at the Medical Imaging Meets NeurIPS workshop (MedNeurIPS). We refer the reader to Appendix D for more details. As for the contributions, Naga Karthik was responsible for carrying out the entire study.

> Karthik, E.N., Kerbrat, A., Labauge, P., Granberg, T., Talbott, J., Reich, D.S., Filippi, M., Bakshi, R., Callot, V., Chandar, S. and Cohen-Adad, J., 2022. Segmentation of Multiple Sclerosis Lesions across Hospitals: Learn Continually or Train from Scratch?. https://arxiv.org/pdf/2210.15091

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

6.1 Introduction

Spinal cord segmentation is relevant for quantifying morphometric changes, such as cord atrophy in multiple sclerosis (MS) [102, 223, 224], compression severity in degenerative cervical myelopathy (DCM) [187, 225], and spared tissue in spinal cord injury (SCI) [184]. 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) [42, 226].

While automatic tools for spinal cord segmentation exist, they have typically been developed in isolated, static environments [17, 122, 174, 227–229]. As a result, these tools rely on different procedures for creating ground truth (GT) masks, model architectures, and training strategies. Chen *et al.* [227] presented an atlas-based topology-preserving method for segmenting scans with different fields of view. Gros *et al.* [17] 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.* [228] 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.* [122] focused on the segmentation of compressed spinal cords with 2D UNets using transfer learning from DeepLabv3 models [230]. Bédard *et al.* [18] introduced `contrast_agnostic`, a 3D model trained on a public dataset of healthy participants [16], 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 [153], 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 [16, 18] 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 [9, 106, 231–234]. One approach to mitigate this variability is to compare them with morphometrics obtained from healthy controls [9, 106, 225, 234, 235]. These normalization techniques assume that the morphometrics of new subjects are computed using the same method as the original normative database [9]. 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 [173], which has been widely adopted by the medical imaging community due to its robustness and generalization to several modalities and neural network architectures [98], 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 [236–238] 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 [222].

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.

3. Validation of the lifelong learning framework to update a normative database of spinal cord morphometric measures [9].

The proposed spinal cord segmentation model and normative database are open-source and integrated into the Spinal Cord Toolbox (SCT) [86], accessible as of v7.0.

6.2 Materials and Methods

6.2.1 Data curation and training protocol

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). The study included 9 different contrasts, namely, T1-weighted (T1w; $n_{\text{vol.}} = 318$), T2-weighted (T2w; $n_{\text{vol.}} = 1377$), T2*-weighted (T2*w; $n_{\text{vol.}} = 499$), diffusion-weighted (DWI; $n_{\text{vol.}} = 243$), gradient-echo sequence with (MT-on; $n_{\text{vol.}} = 248$) and without (GRE-T1w; $n_{\text{vol.}} = 243$) magnetization transfer pulse, phase-sensitive inversion recovery (PSIR; $n_{\text{vol.}} = 333$), short tau inversion recovery (STIR; $n_{\text{vol.}} = 89$), and MP2RAGE UNIT1 ($n_{\text{vol.}} = 103$). 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 6.1 shows the overall summary of the dataset and Table E.1 provides more details on the distribution of image resolutions for each contrast.

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.* [18]. For the newly obtained datasets, we initially performed a quality control (QC) using `sct_qc`, SCT's visual QC tool [226]. Four experienced raters (ENK, SB, JV, JCA) qualitatively assessed the image-GT pairs and

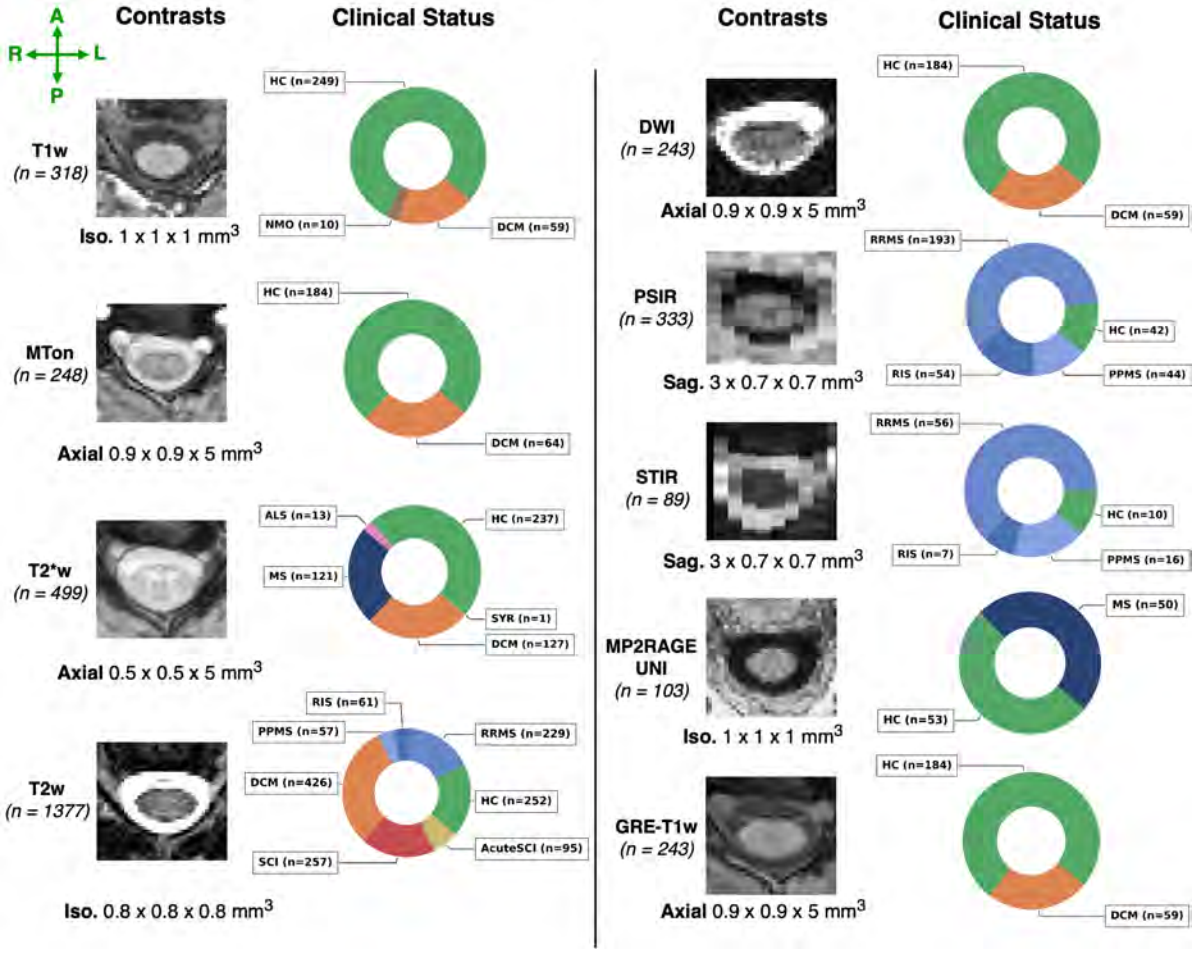

Figure 6.1 Overview of the dataset and image characteristics. Representative axial slices of 9 contrasts and the total of images used for each contrast in brackets, the orientation (axial/sagittal) along with the median resolution of images. The respective doughnut chart illustrates the proportion of 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.

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 [18] and manual corrections. In datasets with severe deformations to the spinal cord anatomy (e.g.,

SCI and DCM), a pathology-specific model, SCIseg [153, 184] 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.

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 [86], SlicerNNUnet, 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 $[0.9 \times 0.7 \times 1]$ 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. 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. $32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 320 \rightarrow 320$). The network was trained with a combination of Dice [134] and cross-entropy losses. At each layer, deep supervision [210] 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.

6.2.2 Lifelong learning for morphometric drift monitoring

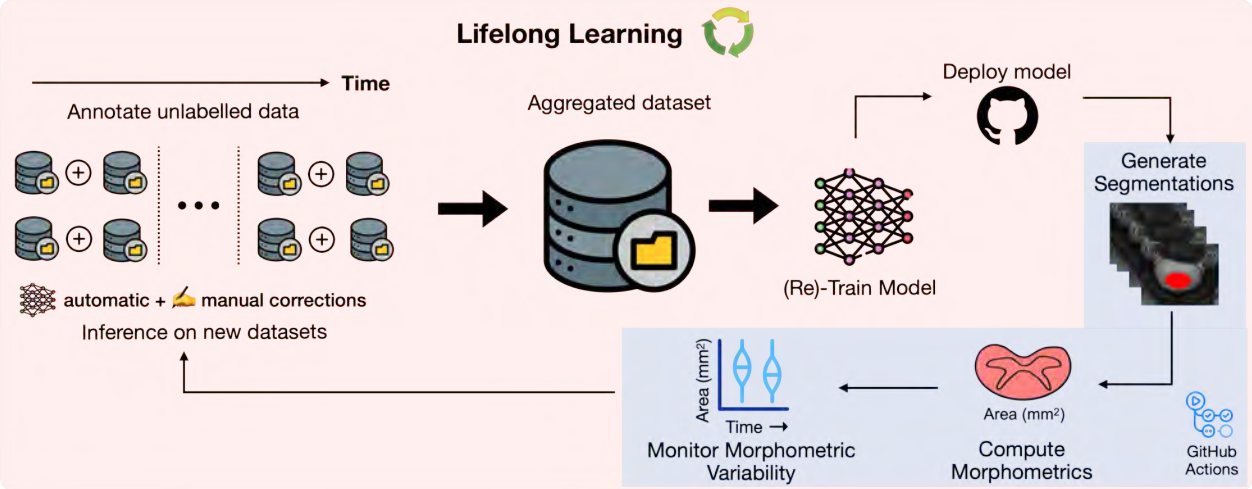

Figure 6.2 Overview of the lifelong learning strategy for continuous training of spinal cord segmentation models. Unlabelled data containing various contrasts and pathologies, gathered from multiple sites worldwide, are segmented automatically with an existing state-of-the-art model and undergo visual 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 [239–241] approach to propose our lifelong learning framework for monitoring morphometric drift across various versions of the model (Figure 6.2). Once the segmentation model is trained, we deploy the model as an official release on GitHub[1]. 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 Algorithm 1 for pseudocode of the workflow). To ensure a fair com-

[1]https://github.com/sct-pipeline/contrast-agnostic-softseg-spinalcord/releases/tag/v3.0

parison with our previous work[2] [18], 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 [18]. 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 [88], as our new models have unrestrained access to all prior data.

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
```

Our choice of using GitHub Actions workflow stems from the ease of accessibility of previous spinal cord segmentation models in SCT [86]. When a new model is released on GitHub, it can be easily downloaded using the command `sct_deepseg spinalcord -install -custom-url`

[2] https://github.com/sct-pipeline/contrast-agnostic-softseg-spinalcord/releases/tag/v2.0

`<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).

6.2.3 Validation protocol

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 set 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.

Qualitative evaluation 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 [242] 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,

namely, `sct_deepseg_sc` [17] (for MS) and SCIsegV2 [153, 184] (for SCI and DCM) using the Dice score, Relative Volume Error (RVE) and Surface Distance, from the ANIMA toolbox [176].

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` [17], and `contrast_agnostic_v2.0` [18].

Ablation study with recursively-generated GT spinal cord masks

As described previously, 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.

Updating the normative database of spinal cord morphometrics

The database of healthy adult morphometrics proposed by Valošek *et al.* [9] included morphometrics measures computed from 203 healthy individuals from the open-access Spine Generic Multi-Subject dataset [16]. These morphometric measures were obtained from segmentations generated with `sct_deepseg_sc` [17], with manual corrections for over/under segmentation errors. As outlined in the Section 6.1, 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 [9].

2. Perform a manual quality control of the spinal cord segmentation masks. Compute 6 morphometric measures (CSA, anteroposterior diameter, transverse diameter, compression ratio, eccentricity and solidity) from the segmentation masks [9].

3. Compute a scaling factor between the morphometric measures derived from different segmentation models, allowing for comparison of morphometric measures across segmentation models.

$$\text{scaling factor} = \frac{<\text{metric}>_{\texttt{contrast\_agnostic\_v3.0}}}{<\text{metric}>_{\texttt{sct\_deepseg\_sc}}}$$

6.3 Results

6.3.1 Evaluation on various contrasts and pathologies

Qualitative comparison of segmentations

Figure 6.3 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 [18].

Figure 6.4 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) and have never been encountered during training. 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.

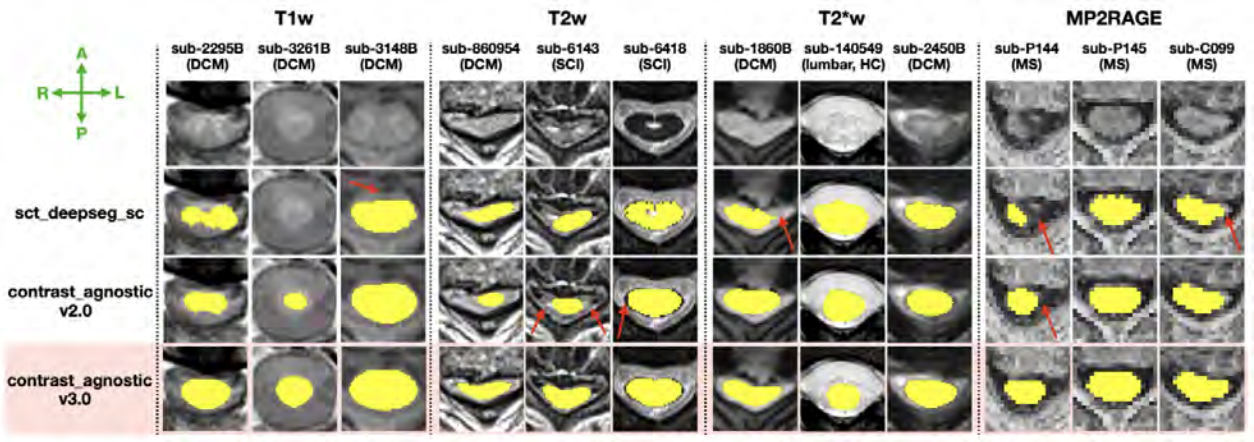

Figure 6.3 Comparison of segmentations between `contrast_agnostic_v3.0` (current version, highlighted), `contrast_agnostic_v2.0` (previous version) and `sct_deepseg_sc` on healthy controls (HC), DCM, SCI and MS patients on the test set (unseen during training). Red arrows show the instances where the previous models fail, particularly under heavy compression (with/without lesions) in sub-860594, sub-6143 and sub-1860B.

Quantitative evaluation on healthy controls and pathologies

Table 6.1 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 6.1A), we then present results for test sets containing T2*w images of MS patients from two sites (Table 6.1B), axial and sagittal T2w scans of DCM patients from two sites (Table 6.1C), and axial and sagittal T2w scans of traumatic SCI (acute, intermediate and chronic phases) from six sites (Table Table 6.1D). 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.

6.3.2 Quantitative evaluation of morphometric drift across model versions

Variability of CSA across contrasts

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

Figure 6.5 shows the CSA STD across six contrasts on the test set of healthy participants

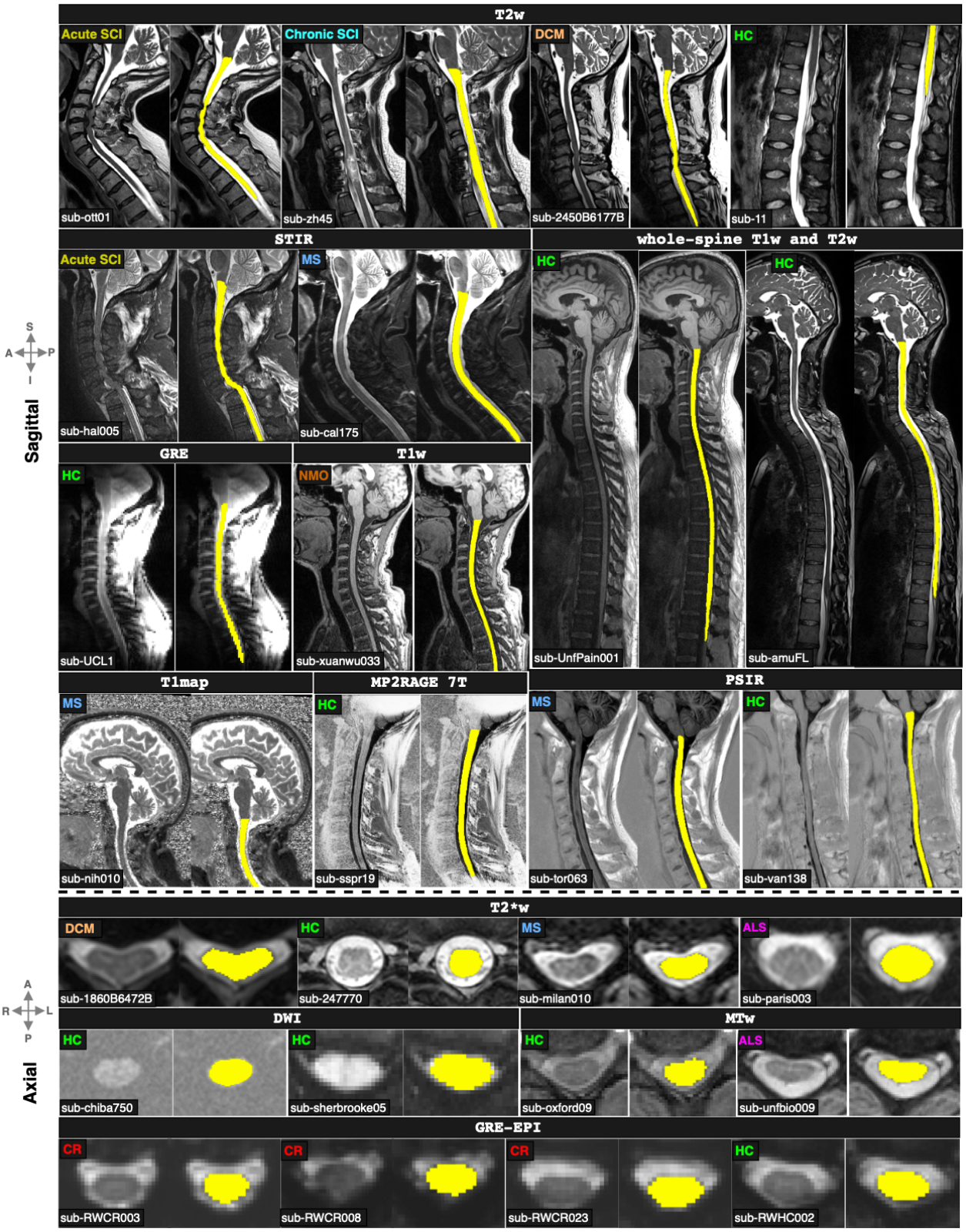

Figure 6.4 Qualitative visualization of the proposed `contrast_agnostic_v3.0` model's segmentations across various contrasts and pathologies on test images from multiple sites. Our model accurately segments compressed spinal cords, severely damaged cords due to injury, and cords with the presence of 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.

Table 6.1 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. Best results are in **bold**.

Methods	**Dice (↑)** Opt. value: 1	**RVE (%)** Opt. value: 0	**ASD (↓)** 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

($n = 49$; $n_{vol} = 294$) of the spine-generic Multi-Subject database [16] between three methods: (i) `sct_deepseg_sc` [17], (ii) `contrast_agnostic_v2.0` [18], 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 E.1.

In Figure 6.6, 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

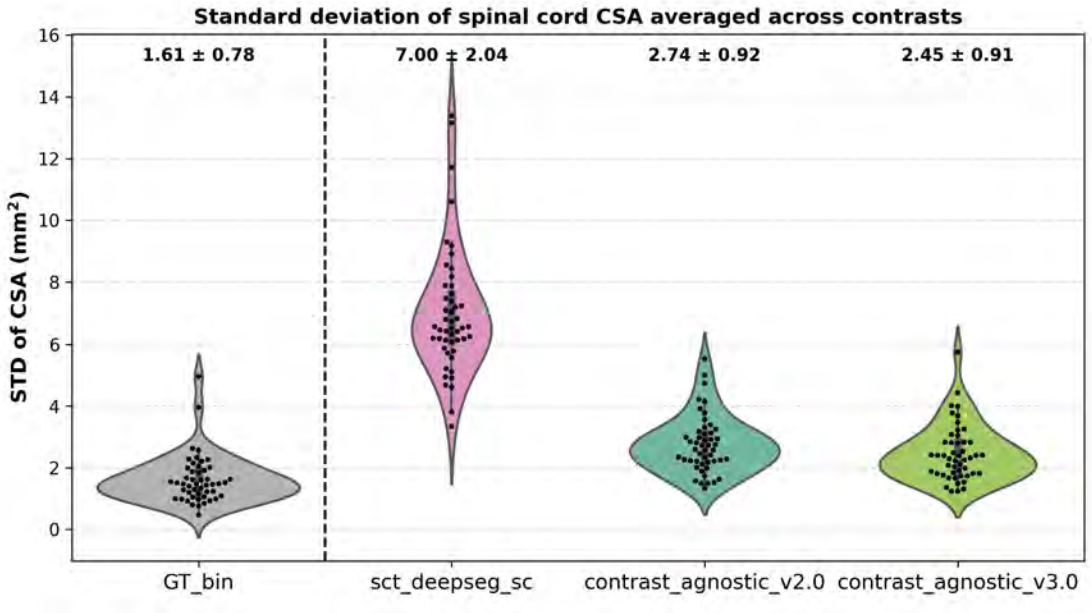

Figure 6.5 CSA variability measured in terms of the standard deviation across 6 contrasts on a test set of healthy participants ($n = 49$). Our proposed model achieved the lowest STD averaged across 6 contrasts (i.e. each point shows the mean of 6 contrasts for the given participant) showing more stability in segmentations across contrasts. The lower the CSA STD across contrasts, the better.

contrasts as `contrast_agnostic_v2.0` was trained only on a healthy participant database.

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 6.7 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

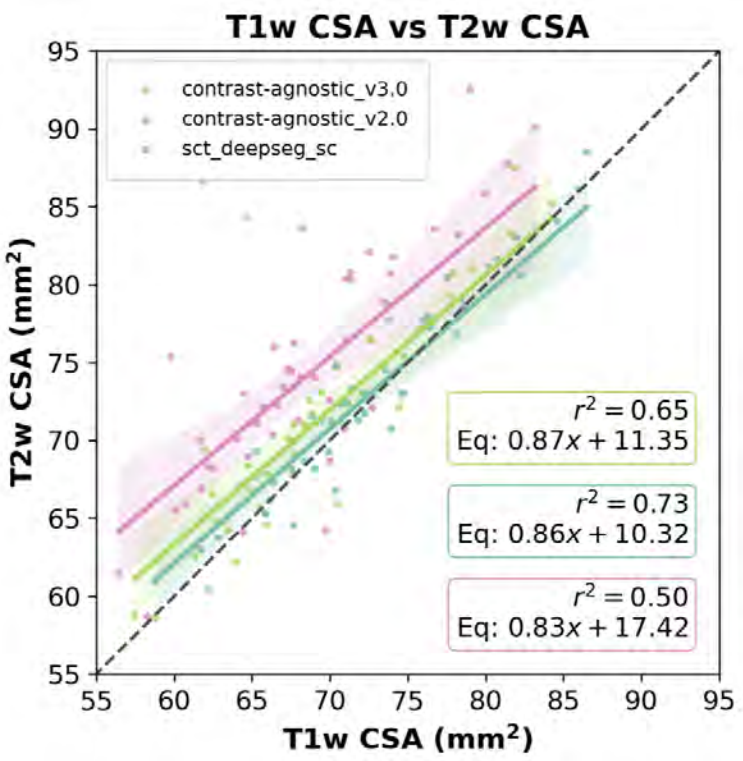

Figure 6.6 Level of agreement between CSA at C2-C3 on T1w and T2w contrasts for `contrast_agnostic_v3.0`, `contrast_agnostic_v2.0` and `sct_deepseg_sc`. Each point represents one participant. The black dashed line represents perfect agreement between the CSA of T1w and T2w contrasts.

masks. In Figure E.2, 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.

Normative database results

Figure 6.8A shows the plots for 6 different morphometric measures computed on 203 healthy participants using two versions of segmentation masks: (i) segmentations from `sct_deepseg_sc` with manual corrections (pink) used in [9] and (ii) segmentations from the proposed model `contrast_agnostic_v3.0` (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 Figure 6.8B. 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

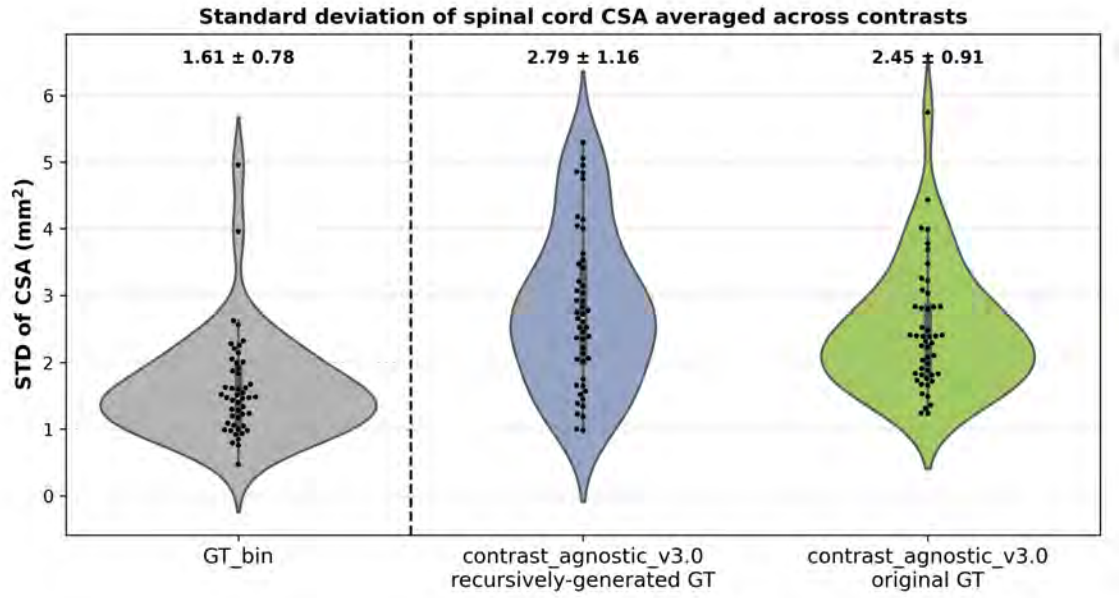

Figure 6.7 Standard deviation of the CSA across 6 contrasts for models trained on: (i) recursively generated GT masks (blue), and (ii) original GT masks (green). Each point shows the mean of 6 contrasts for the given participant. The model trained on noisy labels tends to produce stable segmentations resulting in a lower STD across contrasts. The lower the CSA STD across contrasts, the better.

made open-source[3].

6.4 Discussion and Conclusion

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

[3]https://github.com/spinalcordtoolbox/PAM50-normalized-metrics/releases/tag/r20250321

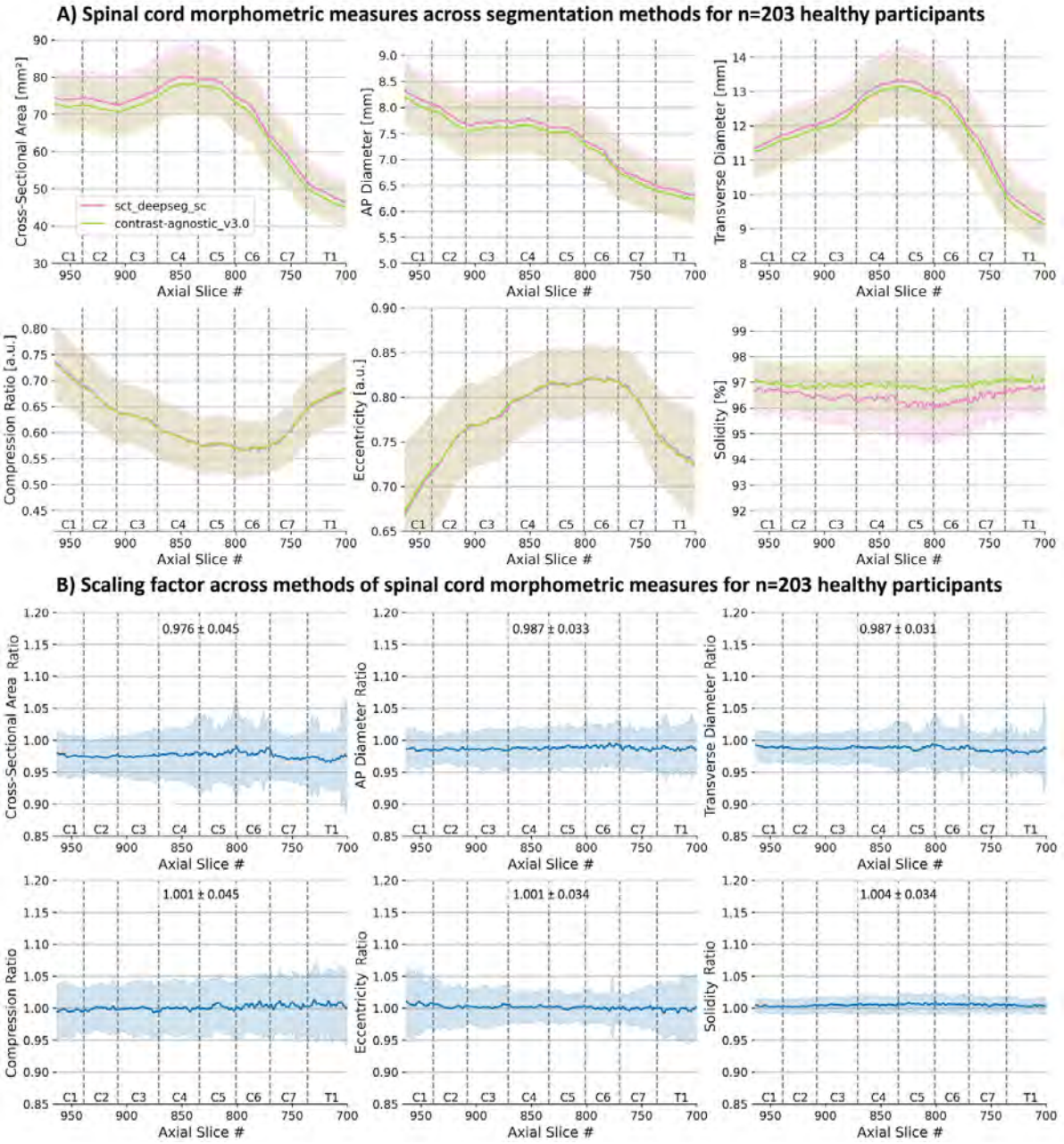

Figure 6.8 (A) Morphometric measures computed on $n = 203$ healthy participants from the Spine Generic Dataset [16] for 6 morphometric measures using 2 different segmentation methods: `sct_deepseg_sc` with manual correction (green) and `contrast_agnostic_v3.0` (orange) with (B) scaling factor between the methods means ± std are displayed. Metrics are shown in the PAM50 space.

update the morphometric measures of a normative database of healthy adults.

6.4.1 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 [243, 244]. However, noise in training data tends to disrupt model training by making the models unintentionally focus on such outliers [93, 245], 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 [17, 18, 153] 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 [246].

6.4.2 Lifelong learning segmentation of the spinal cord

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 6.3 and Figure 6.4, 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 [247]. The competitive performance of the proposed model compared to existing pathology-specific models (Table 6.1) 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.

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 [241, 248, 249]. 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 [237]. 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) [250]. 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 [9, 251]. 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. 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.

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 [236, 237, 250], 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 6.7, 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 [252] has shown that multiple iterations of training on recursively generated data tend to make the model catastrophically forget [88] 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.

Binary vs. soft masks

While training directly on soft masks still achieves the lowest morphometric variability across contrasts [18], 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 [53]. 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.

6.4.3 Application on normative database of morphometrics

Keeping an updated normative morphometrics database is crucial to maintaining lifelong models [9], 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.

6.4.4 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[4][5], 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 - 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 [253] could ensure the continued relevance of the training and test sets for evaluating future models. If the drift between data distribution is large, keeping

[4]https://portal.fli-iam.irisa.fr/ms-multi-spine/
[5]https://ivdm3seg.weebly.com

only a subset of the old data when training new models is recommended.

6.4.5 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.

CHAPTER 7 GENERAL DISCUSSION

In the three previous chapters, we have presented three studies addressing our primary objective of developing open-source, generalizable segmentation tools to better estimate imaging biomarkers. In this chapter, we bring together the contributions from each of these studies and provide some perspectives on automatic spinal cord and lesion segmentation, discuss the clinical impact and present a few possible directions for future research.

7.1 Towards Continuous and Generalizable Spinal Cord and Lesion Segmentation

7.1.1 Human-in-the-loop active learning helps alleviate manual annotation bottlenecks

The lack of large, diverse datasets and the high cost of expert annotations are recurring themes in medical image segmentation tasks [24, 147, 254]. Combining the existing challenges with spinal cord imaging along with the limited prevalence of certain neurological diseases, obtaining large datasets for training robust ML models for spinal cord-specific tasks is challenging. Designing standardized acquisition protocols [42], establishing disease-specific consortiums (CanProCo [255], Praxis (praxisinstitute.org), etc.), and organizing open-source challenges [83–85, 176] present some of the ways to build large-scale and diverse datasets. Gathering diverse datasets solves only half the problem; labeling them presents the more significant challenge. There is a need for automatic tools as several MRI-based clinical biomarkers, especially for disease prognosis and diagnosis and monitoring disease progression, depend on the robust segmentation of the spinal cord segmentation as a key prerequisite [102, 106, 184].

Iterative, semi-automatic approaches in data labeling reduce annotation costs while also scaling to large cohorts [60, 147, 148]. Given a small, expert-annotated dataset for initial training, models can be iteratively improved by bootstrapping automatic predictions with partial human oversight via quality control, thereby reducing the need for expert intervention when labeling large datasets. To this end, we have applied human-in-the-loop active learning as a recurring approach in this thesis to help create large datasets with spinal cord and lesion masks *and*, in the process, develop robust segmentation tools that could aid semi-automatic labeling in future studies. In Chapter 4, we have demonstrated that a three-phase AL approach, by refining intermediate segmentations and iterative training on a diverse

cohort, produced automatically segmented lesion biomarkers showing high agreement with manually-derived measures. Likewise, in Chapter 6, we have utilized the pathology-specific models to generate spinal cord segmentations for a wide variety of contrasts and developed a contrast-agnostic model trained on multisite dataset consisting of 75 sites.

As the barrier to train ML models on large medical datasets decreases with advancements in image acquisition technologies and decreasing computational costs, it is imperative to create high-quality labeled datasets that allow these models generalize to unseen data distributions, ultimately translating to clinical workflows.

7.1.2 Importance of simple yet rigorously-validated architectures

The existence of benchmarking datasets such as BraTS [70] and Decathlon [256] has resulted in the development of numerous segmentation models spanning various classes of model architectures [24, 25, 257]. Therefore, with such variety in model architectures comes the issue of actual *utility* of these models. Most methods tend to be developed in highly specific computational environments without rigorous validation on unseen, out-of-distribution data [98, 141, 142]. Furthermore, it is not difficult to notice that most approaches report merely incremental improvements over previous methods (sometimes even without error intervals! [142]). Therefore, this begs the question, *do increasingly complex architectures offer substantial advantages over proven models such as nnUNet, or do they only provide marginal improvements at a larger computational cost?* Benchmarking studies have reported that CNN models achieve both speed and accuracy over transformer and SSM-based models, making it a tradeoff between efficiency and computational cost [129, 131]. This is concerning because researchers routinely face the dilemma of spending months tinkering with fancy new architectures that do not guarantee significant improvements on their task-of-interest, or, going with robust, time-tested approaches and risk rejection as they are not "novel" enough.

We have focused on the robust, widely-adopted, CNN-based architecture for developing the segmentation tools in this thesis. With the rise of large language models, there has a been visible push towards adopting transformer-based architectures [131]. However, naive transformer models are still compute-hungry, requiring large amounts of GPU memory for the quadratic-complexity self-attention mechanism and typically require larger training datasets [131]. Moreover, as seen in Chapter 5, large images patches are required to model the context around the spinal cord, which, when using transformers, quickly runs into out-of-memory errors as they typically utilize small patches [257]. Lastly, studies such as [98, 129, 131] have pointed out that despite the hype towards novel architectures, CNN-based architectures, when appropriately-tuned for a given segmentation task, outperform most state-of-the-art

methods under a fraction of the total training time and computational resources, further reinforcing our decision to use simple CNNs.

7.1.3 Need for clinically-oriented measures beyond pure segmentation metrics

We are witnessing an exponential rise in the various ways in which ML models are applied for medical imaging tasks [1, 148]. Yet, only a tiny fraction of these models are translated into clinical workflows [14]. Two of the key issues in this failure to translate to clinical practice are: (i) inappropriate choice of validation metrics that do not accurately evaluate the underlying task, and (ii) over-reliance of chosen quantitative metrics as *the* measure of superiority over other methods. For instance, considering how the most-commonly used measure, Dice similarity score, provides a false impression when evaluating small structures (Figure 2.15). Instead of relying on one or two metrics, it is critical to use multiple metrics across different families (overlap-based, volume-based, and distance-based) to account for their complementary properties.

While selecting appropriate metrics across different classes is beneficial, it is not a complete solution, as metric values can be influenced when used as loss functions for neural network training. Hence, we went a step beyond in this thesis. Specifically, with each automatic tool presented, we have identified appropriate clinically-targeted measures to evaluate our proposed approach in addition to reporting the standard metrics quantifying segmentation accuracy. For instance, in Chapter 4, we have used intramedullary lesion length, total lesion volume, and tissue bridges as surrogate measures to evaluate the clinical utility of SCIseg compared to other approaches. Likewise, in Chapters 5 and 6, we have defined the variability in the spinal cord cross-sectional area (CSA) across contrasts as a surrogate clinical metric to compare the performance of various approaches. It is important to note that in tasks where the object-of-interest is large with respect to the background (*e.g.* spinal cord), competing approaches often perform similarly with a few percentage points of difference (*e.g.* see Table 5.2, Figure 4.4, and Table 6.1). In such cases, it becomes imperative to define proxy measures (in our case, quantifying the *variability* of spinal cord CSA) to observe meaningful differences between competing methods (Figures 5.6, 5.7 and 5.9).

7.1.4 Facilitating continuous development of segmentation models

Similar to how software evolves in its life cycle, ML models too will continue to develop with newer models becoming capable of solving diverse range of tasks. In the context of large language models, continued pretraining is one of the approaches used to steer models towards new or out-of-distribution domains of knowledge [258, 259]. Likewise, segmentation models

trained on a static, "timestamped" version of a dataset will continue to evolve and refine their segmentation capabilities as datasets are updated with new contrasts and pathologies added over time. Therefore, it is critical to design segmentation frameworks that allow for *continuous* training in the background once the model is deployed. More importantly, as medical data are gathered from heterogenous sources including different scanners and acquisition protocols, monitoring for performance drift over subsequent versions is crucial. With enough iterations, such a lifelong learning framework can enable segmentation models to handle out-of-distribution data effectively.

To this end, we presented a post-deployment lifelong learning scenario for spinal cord segmentation as a proof-of-concept in Chapter 6. Borrowing from standard MLOps practices, we set up a GitHub Actions workflow to automatically monitor for performance drifts (specifically, the morphometric variability in spinal cord CSA) across various model versions. Initially, human intervention was required to refine automatic predictions for further training, preventing the framework from being fully autonomous. Yet, this manual step may no longer be necessary as the models continuously improve.

7.2 Prospect of Application and Clinical Impact

Manual estimation of imaging biomarkers is a bottleneck that introduces rater bias and limits monocentric studies from scaling to large, multi-site cohorts. The existence of automatic tools developed using diverse, multi-site datasets has the potential to aid future studies by reducing rater bias and result in objective estimation of imaging biomarkers. To this end, the reliability of SCIseg predictions for measuring midsagittal tissue bridges was evaluated in a recent study [151], concluding that intra-class correlation coefficient between manual and automated measures were excellent. Furthermore, Schading-Sassenhausen *et al.* [152] studied the relation between spinal tract damage and the development of spastic muscle tone in SCI patients using automatic spinal cord and lesion masks. Oliva *et al.* [260] used our contrast-agnostic spinal cord segmentation tool to obtain segmentation masks from motion-corrected functional images to study hand function using simultaneous brain-spinal cord functional MRI. Likewise, Muhammad *et al.* [123] used automatic cord segmentations to characterize morphometrics of the compressed spinal cord in DCM patients. Together, these studies demonstrate the clinical utility of the segmentation tools presented in this thesis.

7.3 Avenues for Future Research

Longitudinal analysis The robust segmentation tools for lesions and the spinal cord can be employed for longitudinal monitoring of imaging biomarkers in SCI. For instance, measuring the volume and length of the lesions, and the remaining tissue bridges surrounding the lesion could provide an objective, unbiased approach to guiding rehabilitation decision making and stratifying patients into homogeneous subgroups of recovery in clinical trials.

Likewise, monitoring cord atrophy could enhance our pathophysiological understanding of various neurological diseases. In MS, cervical spinal cord atrophy correlates with disability and can predict conversion to progressive MS. Monitoring this atrophy using the contrast-agnostic spinal cord segmentation tool can help in tracking disease progression, assess the impact of interventions, and identifying patients at a higher risk to improve patient selection for clinical trials.

Model Zoo / Weight-space learning For solving a particular task using DL, it is common to train hundreds (or even thousands) of models with different hyperparameter combinations. This experimentation phase results in several checkpoints (*i.e.* `.pt` files) containing model parameters/weights required to solve the task up to a certain extent (assuming it is trained until convergence). Out of this bag of checkpoints, one model whose parameters obtain the best validation accuracy is chosen and the rest are discarded. Consider a scenario where these checkpoints are *not* discarded. Given the recent surge in diffusion models for image generation [261, 262], *what if we train a diffusion model on the pretrained checkpoints to generate the optimal set of weights used to solve a particular task*? The idea of using generative models on model checkpoints was first proposed by Peebles *et al.* [263], where, using the checkpoints as input, a diffusion model was trained to generate optimized network parameters. At test time, by providing a loss value and a set of randomly-initialized parameters, the diffusion model generates a set of optimal parameters, essentially solving the downstream classification task in one shot.

Currently, SCT hosts several models, each for solving a particular segmentation task (*e.g.* spinal cord, lesions, gray matter, canal, etc.) using a similar nnUNet-based architecture. By gathering the weights of the deployed models *and* the weights of all the models in the experimentation phase, one could create a dataset of nnUNet checkpoints. Then, using the previous approach, a diffusion model could be trained to generate the optimal weights for *any* of the tasks. The main advantage here is that instead of having N models for N tasks, we could have one model solving N tasks as it has been trained on checkpoints of all these tasks. Proof-of-concept results from the original paper were reported on MNIST digit classification

pretrained with thousands of checkpoints. Its scalability of larger models and more complex tasks could be a critical limitation that remains to be explored.

Leveraging multiple modalities All the methods presented in this thesis are *unimodal*, that is, the models have been trained with image data (spinal cord MRI scans) only. Recent research has shown that models fusing data from multiple modalities perform better than unimodal models on average [264,265]. Text, in the form of natural language, is the one of the commonly used modality in medicine. Unlike X-ray datasets containing paired text-image data with radiographic reports describing the contents of the image, gathering text data in MRI datasets is challenging. Researchers typically use predefined templates or large language models (GPT-4o) for generating text [266]. In addition to textual input, tabular data is another common modality combined with images. Existing studies have used clinical tables to train multimodal models and have reported improvments over unimodal ones [267,268]. Future works could consider utilizing subject-specific metadata from `participants.tsv` file in BIDS-structured datasets as additional inputs during training. For example, the inputs could be conditioned on the subject age/sex/clinical status to encode subject-specific information.

Augmenting SCT with AI agents We have presented two tools, contrast-agnostic spinal cord segmentation [18] and lesion segmentation in SCI patients [153], both accessible via the SCT package. Likewise, segmentation tools for lesions [247], rootlets [269] and other spinal structures [270] have been developed. These DL-based tools complement the existing image processing functionalities of SCT including registration (`sct_register_multimodal`), motion correction (`sct_dmri_moco`), segmentation analysis (`sct_analyze_lesion`), etc. Notably, the aforementioned tools are accessible to the user via command line interface. Alternatively, with the success of large language models, a system can use the language models as "agents" that can use external tools to perform tasks beyond text generation. Through a unified natural language interface, this approach coordinates multiple specialized models, enabling them to collaborate on complex tasks that exceed the capabilities of any single model [271,272]. For example, currently, if a user wants segment the spinal cord, they would have to run: `sct_deepseg spinalcord -i <path/to/image>.nii.gz -o <path/to/output>.nii.gz`. As SCT evolves, the command used to run the segmentation algorithm could change. In the scenario where SCT is augmented with agentic capabilities, a simple prompt: `Output the spinal cord segmentation mask for this image <upload image>`, could perform this task in three steps: (i) trigger the agent that fetches the spinal cord segmentation tool, (ii) runs inference with the model, and (iii) outputs the mask. While the example only highlights spinal cord segmentation, such prompts could be used to segment lesions or generate lesion

statistics. As the tools within SCT get more robust, having an agent interact with all the available tools simplifies the maintenance while also making it more user-friendly.

CHAPTER 8 CONCLUSION

In this thesis, we have presented tools for the automatic segmentation of the spinal cord and lesions across a wide spectrum of contrasts and pathologies for improved estimation of imaging biomarkers. We have explored human-in-the-loop active learning as a means to iteratively annotate and train models on large, diverse training cohorts and developed a lifelong learning framework facilitating continuous development of spinal cord segmentation models. Segmentation models in medical imaging will continue to be developed in the future. This thesis advocates for better evaluation of such models by identifying the clinically-targeted quantitative measures going beyond pure segmentation metrics.

REFERENCES

[1] J. Ma *et al.*, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[2] F. Martini, W. Ober, and J. Nath, *Visual Anatomy & Physiology*. Pearson Education, 2012. [Online]. Available: https://books.google.ca/books?id=Zs_cAgAAQBAJ

[3] W. Commons. (2010) Diagrammatic transverse section of the medulla spinalis and its membranes. [Online]. Available: https://commons.wikimedia.org/wiki/File:Gray770-en.svg

[4] A. Grayev, "Functional anatomy of the spinal cord," *Radiologic Clinics*, vol. 62, no. 2, pp. 263–272, 2024.

[5] M. Bath. (2025) The spinal cord. [Online]. Available: https://teachmeanatomy.info/back/nerves/spinal-cord/

[6] A.-J. Fordham *et al.*, "Differentiating glioblastomas from solitary brain metastases: An update on the current literature of advanced imaging modalities," *Cancers*, vol. 13, p. 2960, 06 2021.

[7] M. Jenkinson and M. Chappell, *Introduction to neuroimaging analysis*. Oxford University Press, 2018.

[8] S. C. Toolbox. (2017) The partial volume effect. [Online]. Available: https://spinalcordtoolbox.com/stable/user_section/tutorials/atlas-based-analysis/partial-volume-effect.html

[9] J. Valošek *et al.*, "A database of the healthy human spinal cord morphometry in the pam50 template space," *Imaging Neuroscience*, vol. 2, pp. 1–15, 2024.

[10] A. Cohen-Gadol. (2024) Spinal cord injury: What the patient needs to know. [Online]. Available: https://www.aaroncohen-gadol.com/en/patients/spinal-cord-injury/overview

[11] M. Seif *et al.*, "Guidelines for the conduct of clinical trials in spinal cord injury: Neuroimaging biomarkers," *Spinal Cord*, vol. 57, no. 9, pp. 717–728, 2019.

[12] K. Mileski. (2024) Degenerative cervical myelopathy treatment. [Online]. Available: https://propelphysiotherapy.com/spinal-cord-injury/degenerative-cervical-myelopathy-treatment/

[13] M. Clinic. (2024) Multiple sclerosis. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/multiple-sclerosis/symptoms-causes/syc-20350269

[14] L. Maier-Hein *et al.*, "Metrics reloaded: recommendations for image analysis validation," *Nature methods*, vol. 21, no. 2, pp. 195–212, 2024.

[15] C. Gros, A. Lemay, and J. Cohen-Adad, "Softseg: Advantages of soft versus binary training for image segmentation," *Medical Image Analysis*, vol. 71, p. 102038, 2021.

[16] J. Cohen-Adad *et al.*, "Open-access quantitative MRI data of the spinal cord and reproducibility across participants, sites and manufacturers," *Scientific Data*, vol. 8, no. 1, p. 219, Aug. 2021.

[17] C. Gros *et al.*, "Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks," *Neuroimage*, vol. 184, pp. 901–915, Jan. 2019.

[18] S. Bédard *et al.*, "Towards contrast-agnostic soft segmentation of the spinal cord," *Medical Image Analysis*, p. 103473, 2025.

[19] H.-D. Cheng *et al.*, "Color image segmentation: advances and prospects," *Pattern recognition*, vol. 34, no. 12, pp. 2259–2281, 2001.

[20] Y. Guo *et al.*, "A review of semantic segmentation using deep neural networks," *International journal of multimedia information retrieval*, vol. 7, pp. 87–93, 2018.

[21] S. Ghosh *et al.*, "Understanding deep learning techniques for image segmentation," *ACM computing surveys (CSUR)*, vol. 52, no. 4, pp. 1–35, 2019.

[22] S. Minaee *et al.*, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[23] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," *Annual review of biomedical engineering*, vol. 2, no. 1, pp. 315–337, 2000.

[24] M. E. Rayed *et al.*, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in Medicine Unlocked*, p. 101504, 2024.

[25] R. Azad *et al.*, "Medical image segmentation review: The success of u-net," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[26] A. Carass *et al.*, "Longitudinal multiple sclerosis lesion segmentation: resource and challenge," *NeuroImage*, vol. 148, pp. 77–102, 2017.

[27] L. Pei *et al.*, "Longitudinal brain tumor segmentation prediction in mri using feature and label fusion," *Biomedical signal processing and control*, vol. 55, p. 101648, 2020.

[28] M. Diaz-Hurtado *et al.*, "Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review," *Neuroradiology*, vol. 64, no. 11, pp. 2103–2117, 2022.

[29] M. R. Rokuss *et al.*, "Longitudinal segmentation of ms lesions via temporal difference weighting," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 64–74.

[30] S. K. Warfield, F. A. Jolesz, and R. Kikinis, "Real-time image segmentation for image-guided surgery," in *SC'98: Proceedings of the 1998 ACM/IEEE Conference on Supercomputing*. IEEE, 1998, pp. 42–42.

[31] A. Madani *et al.*, "Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy," *Annals of surgery*, vol. 276, no. 2, pp. 363–369, 2022.

[32] J. Prince and J. Links, *Medical Imaging Signals and Systems*, ser. Pearson Prentice Hall bioengineering. Pearson Prentice Hall, 2006. [Online]. Available: https://books.google.ca/books?id=IPm8QgAACAAJ

[33] F. Bruno *et al.*, "Advanced magnetic resonance imaging (mri) of soft tissue tumors: techniques and applications," *La radiologia medica*, vol. 124, pp. 243–252, 2019.

[34] P. Freund *et al.*, "Mri in traumatic spinal cord injury: from clinical assessment to neuroimaging biomarkers," *The Lancet Neurology*, vol. 18, no. 12, pp. 1123–1135, 2019.

[35] F. Barkhof and K. K. Koeller, "Demyelinating diseases of the cns (brain and spine)," *Diseases of the brain, head and neck, spine 2020–2023: diagnostic imaging*, pp. 165–176, 2020.

[36] M. Etemadifar *et al.*, "Mri signs of cns demyelinating diseases," *Multiple Sclerosis and Related Disorders*, vol. 47, p. 102665, 2021.

[37] F. Nelson *et al.*, "Improved identification of intracortical lesions in multiple sclerosis with phase-sensitive inversion recovery in combination with fast double inversion recovery mr imaging," *American Journal of Neuroradiology*, vol. 28, no. 9, pp. 1645–1649, 2007. [Online]. Available: https://www.ajnr.org/content/28/9/1645

[38] S. Peters *et al.*, "Detection of spinal cord multiple sclerosis lesions using a 3d-psir sequence at 1.5 t," *Clinical Neuroradiology*, vol. 34, no. 2, pp. 403–410, 2024.

[39] N. Kinany *et al.*, "Spinal cord fmri: A new window into the central nervous system," *The Neuroscientist*, vol. 29, no. 6, pp. 715–731, 2023.

[40] R. L. Barry *et al.*, "Spinal cord mri at 7t," *Neuroimage*, vol. 168, pp. 437–451, 2018.

[41] N. Kinany *et al.*, "Dynamic functional connectivity of resting-state spinal cord fmri reveals fine-grained intrinsic architecture," *Neuron*, vol. 108, no. 3, pp. 424–435, 2020.

[42] J. Cohen-Adad *et al.*, "Generic acquisition protocol for quantitative mri of the spinal cord," *Nature protocols*, vol. 16, no. 10, pp. 4611–4632, 2021.

[43] P. W. Stroman *et al.*, "The current state-of-the-art of spinal cord imaging: methods," *Neuroimage*, vol. 84, pp. 1070–1081, 2014.

[44] M. M. El Mendili *et al.*, "Spinal cord imaging in amyotrophic lateral sclerosis: historical concepts—novel techniques," *Frontiers in neurology*, vol. 10, p. 350, 2019.

[45] E. N. Marieb and K. Hoehn, *Human anatomy & physiology*. Pearson education, 2007.

[46] J. Finsterbusch, "B0 inhomogeneity and shimming," in *Quantitative MRI of the spinal cord*. Elsevier, 2014, pp. 68–90.

[47] D. Norman *et al.*, "Magnetic resonance imaging of the spinal cord and canal: potentials and limitations," *American journal of roentgenology*, vol. 141, no. 6, pp. 1147–1152, 1983.

[48] M. A. G. Ballester, A. P. Zisserman, and M. Brady, "Estimation of the partial volume effect in mri," *Medical Image Analysis*, vol. 6, no. 4, pp. 389–405, 2002.

[49] J. Tohka, "Partial volume effect modeling for segmentation and tissue classification of brain magnetic resonance images: A review," *World journal of radiology*, vol. 6, no. 11, p. 855, 2014.

[50] S. Lévy *et al.*, "White matter atlas of the human spinal cord with estimation of partial volume effect," *Neuroimage*, vol. 119, pp. 262–271, 2015.

[51] B. Billot *et al.*, "Partial volume segmentation of brain mri scans of any resolution and contrast," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23*. Springer, 2020, pp. 177–187.

[52] O. Vincent, C. Gros, and J. Cohen-Adad, "Impact of individual rater style on deep learning uncertainty in medical imaging segmentation," *arXiv preprint arXiv:2105.02197*, 2021.

[53] A. Lemay *et al.*, "Label fusion and training methods for reliable representation of inter-rater uncertainty," *arXiv preprint arXiv:2202.07550*, 2022.

[54] B. Nichyporuk *et al.*, "Rethinking generalization: The impact of annotation style on medical image segmentation," *arXiv preprint arXiv:2210.17398*, 2022.

[55] R. Walsh *et al.*, "Expert variability and deep learning performance in spinal cord lesion segmentation for multiple sclerosis patients," in *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2023, pp. 463–470.

[56] R. M. Monarch, *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

[57] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, 2021.

[58] E. Mosqueira-Rey *et al.*, "Human-in-the-loop machine learning: a state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, 2023.

[59] P. Ren *et al.*, "A survey of deep active learning," *ACM computing surveys (CSUR)*, vol. 54, no. 9, pp. 1–40, 2021.

[60] S. Rahimi *et al.*, "Addressing the exorbitant cost of labeling medical images with active learning," in *International Conference on Machine Learning in Medical Imaging and Analysis*, June 2021.

[61] C. Tan *et al.*, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*. Springer, 2018, pp. 270–279.

[62] S. Motiian *et al.*, "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5715–5725.

[63] Q. Dou *et al.*, "Domain generalization via model-agnostic learning of semantic features," in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/2974788b53f73e7950e8aa49f3a306db-Paper.pdf

[64] H. Li *et al.*, "Domain generalization for medical imaging classification with linear-dependency regularization," *Advances in neural information processing systems*, vol. 33, pp. 3118–3129, 2020.

[65] C. Sendra-Balcells *et al.*, "Domain generalization in deep learning for contrast-enhanced imaging," *Computers in Biology and Medicine*, vol. 149, p. 106052, 2022.

[66] T. Tu *et al.*, "Towards generalist biomedical ai," *Nejm Ai*, vol. 1, no. 3, p. AIoa2300138, 2024.

[67] L. Zhang *et al.*, "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2531–2540, 2020.

[68] S. Zixian *et al.*, "Rethinking data augmentation for single-source domain generalization in medical image segmentation," in *AAAI*, 2023.

[69] B. Billot *et al.*, "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining," *Medical Image Analysis*, vol. 86, p. 102789, May 2023.

[70] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[71] M. Ghaffari, A. Sowmya, and R. Oliver, "Automated brain tumor segmentation using multimodal brain scans: a survey based on models submitted to the brats 2012–2018 challenges," *IEEE reviews in biomedical engineering*, vol. 13, pp. 156–168, 2019.

[72] U. Baid *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[73] M. C. de Verdier *et al.*, "The 2024 brain tumor segmentation (brats) challenge: glioma segmentation on post-treatment mri," *arXiv preprint arXiv:2405.18368*, 2024.

[74] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[75] R. W. Cox, "Afni: software for analysis and visualization of functional magnetic resonance neuroimages," *Computers and Biomedical research*, vol. 29, no. 3, pp. 162–173, 1996.

[76] W. D. Penny *et al.*, *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[77] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[78] M. Jenkinson *et al.*, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[79] O. Commowick *et al.*, "Multiple sclerosis lesions segmentation from multiple experts: The miccai 2016 challenge dataset," *Neuroimage*, vol. 244, p. 118589, 2021.

[80] ADNI. (2025) The ms-multi-spine miccai 2025 challenge. [Online]. Available: https://adni.loni.usc.edu

[81] K. Marek *et al.*, "The parkinson's progression markers initiative (ppmi)–establishing a pd biomarker cohort," *Annals of clinical and translational neurology*, vol. 5, no. 12, pp. 1460–1477, 2018.

[82] M. S. Byun *et al.*, "Korean brain aging study for the early diagnosis and prediction of alzheimer's disease: methodology and baseline sample characteristics," *Psychiatry investigation*, vol. 14, no. 6, p. 851, 2017.

[83] F. Prados *et al.*, "Spinal cord grey matter segmentation challenge," *Neuroimage*, vol. 152, pp. 312–329, 2017.

[84] MS-Multi-Spine Team. (2004) Alzheimer's disease neuroimaging initiative. [Online]. Available: https://portal.fli-iam.irisa.fr/ms-multi-spine/

[85] J. W. van der Graaf *et al.*, "Lumbar spine segmentation in mr images: a dataset and a public benchmark," *Scientific Data*, vol. 11, no. 1, p. 264, 2024.

[86] B. De Leener *et al.*, "Sct: Spinal cord toolbox, an open-source software for processing spinal cord mri data," *Neuroimage*, vol. 145, pp. 24–43, 2017.

[87] S. Thrun, "Lifelong learning algorithms," in *Learning to learn*. Springer, 1998, pp. 181–209.

[88] S. Sodhani *et al.*, "An introduction to lifelong supervised learning," *arXiv preprint arXiv:2207.04354*, 2022.

[89] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*. Elsevier, 1989, vol. 24, pp. 109–165.

[90] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[91] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[92] W. R. Crum, O. Camara, and D. L. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE transactions on medical imaging*, vol. 25, no. 11, pp. 1451–1461, 2006.

[93] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC medical imaging*, vol. 15, pp. 1–28, 2015.

[94] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, no. 1, p. 210, 2022.

[95] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[96] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[97] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[98] F. Isensee *et al.*, "nnu-net revisited: A call for rigorous validation in 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 488–498.

[99] N. Cawley *et al.*, "Spinal cord atrophy as a primary outcome measure in phase II trials of progressive multiple sclerosis," *Multiple Sclerosis*, vol. 24, no. 7, pp. 932–941, Jun. 2018.

[100] A. C. Smith *et al.*, "Lateral corticospinal tract damage correlates with motor output in incomplete spinal cord injury," *Archives of Physical Medicine and Rehabilitation*, vol. 99, no. 4, pp. 660–666, Apr. 2018.

[101] D. Pfyffer *et al.*, "Tissue bridges predict recovery after traumatic and ischemic thoracic spinal cord injury," *Neurology*, vol. 93, no. 16, pp. e1550–e1560, 2019.

[102] P. Bautin and J. Cohen-Adad, "Minimum detectable spinal cord atrophy with automatic segmentation: Investigations using an open-access dataset of healthy participants," *NeuroImage: Clinical*, vol. 32, p. 102849, 2021.

[103] A. C. Smith *et al.*, "Axial MRI biomarkers of spinal cord damage to predict future walking and motor function: a retrospective study," *Spinal Cord*, vol. 59, no. 6, pp. 693–699, Jun. 2021.

[104] R. Bakshi *et al.*, "Measurement of brain and spinal cord atrophy by magnetic resonance imaging as a tool to monitor multiple sclerosis," *Journal of Neuroimaging*, vol. 15, pp. 30S–45S, 2005.

[105] C. Casserly *et al.*, "Spinal cord atrophy in multiple sclerosis: A systematic review and meta-analysis," *Journal of Neuroimaging*, vol. 28, no. 6, pp. 556–586, 2018.

[106] S. Bédard *et al.*, "Normalizing spinal cord compression measures in degenerative cervical myelopathy," *The Spine Journal*, 2025.

[107] A. Birenbaum and H. Greenspan, "Multi-view longitudinal cnn for multiple sclerosis lesion segmentation," *Engineering Applications of Artificial Intelligence*, vol. 65, pp. 111–118, 2017.

[108] J. Krüger *et al.*, "Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3d convolutional neural networks," *NeuroImage: Clinical*, vol. 28, p. 102445, 2020.

[109] D. Pfyffer *et al.*, "Predictive value of midsagittal tissue bridges on functional recovery after spinal cord injury," *Neurorehabilitation and neural repair*, vol. 35, no. 1, pp. 33–43, 2021.

[110] L. Telano and S. Baker, "Physiology, cerebral spinal fluid," in *StatPearls*, 2025th ed. Treasure Island (FL): StatPearls Publishing, October 2018, updated 2023 Jul 4.

[111] D. Purves *et al.*, "The internal anatomy of the spinal cord," in *Neuroscience. 2nd edition.* Sinauer Associates, 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK11008/

[112] S. D. Serai *et al.*, "Components of a magnetic resonance imaging system and their relationship to safety and image quality," *Pediatric radiology*, vol. 51, pp. 716–723, 2021.

[113] D. A. Feinberg *et al.*, "Next-generation mri scanner designed for ultra-high-resolution human brain imaging at 7 tesla," *Nature methods*, vol. 20, no. 12, pp. 2048–2057, 2023.

[114] G. David *et al.*, "Longitudinal changes of spinal cord grey and white matter following spinal cord injury," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 92, no. 11, pp. 1222–1230, 2021.

[115] C. S. Ahuja *et al.*, "Traumatic spinal cord injury," *Nature reviews Disease primers*, vol. 3, no. 1, pp. 1–21, 2017.

[116] A. Alizadeh, S. M. Dyck, and S. Karimi-Abdolrezaee, "Traumatic spinal cord injury: an overview of pathophysiology, models and acute injury mechanisms," *Frontiers in neurology*, vol. 10, p. 282, 2019.

[117] P. Freund *et al.*, "Mri investigation of the sensorimotor cortex and the corticospinal tract after acute spinal cord injury: a prospective longitudinal study," *The Lancet Neurology*, vol. 12, no. 9, pp. 873–881, 2013.

[118] I. Jure and F. Labombarda, "Spinal cord injury drives chronic brain changes," *Neural Regeneration Research*, vol. 12, no. 7, pp. 1044–1047, 2017.

[119] L. M. Shah and J. S. Ross, "Imaging of spine trauma," *Neurosurgery*, vol. 79, no. 5, pp. 626–642, 2016.

[120] J. H. Badhiwala *et al.*, "Degenerative cervical myelopathy—update and future directions," *Nature Reviews Neurology*, vol. 16, no. 2, pp. 108–124, 2020.

[121] G. David *et al.*, "Traumatic and nontraumatic spinal cord injury: pathological insights from neuroimaging," *Nature Reviews Neurology*, vol. 15, no. 12, pp. 718–731, 2019.

[122] K. Nozawa *et al.*, "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*, vol. 18, no. 1, pp. 45–54, 2023.

[123] F. Muhammad *et al.*, "Cervical spinal cord morphometrics in degenerative cervical myelopathy: quantification using semi-automated normalized technique and correlation

with neurological dysfunctions," *The Spine Journal*, vol. 24, no. 11, pp. 2045–2057, 2024.

[124] M. M. Goldenberg, "Multiple sclerosis review," *Pharmacy and therapeutics*, vol. 37, no. 3, p. 175, 2012.

[125] F. Cermelli *et al.*, "Modeling the background for incremental learning in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9233–9242.

[126] P. Chlap *et al.*, "A review of medical image data augmentation techniques for deep learning applications," *Journal of medical imaging and radiation oncology*, vol. 65, no. 5, pp. 545–563, 2021.

[127] D. Azizov *et al.*, "A decade of deep learning: A survey on the magnificent seven," *arXiv preprint arXiv:2412.16188*, 2024.

[128] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[129] D. Gut *et al.*, "Benchmarking of deep architectures for segmentation of medical images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3231–3241, 2022.

[130] P. R. Bassi *et al.*, "Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 15 184–15 201, 2024.

[131] P. M. Kazaj *et al.*, "From claims to evidence: A unified framework and critical analysis of cnn vs. transformer vs. mamba in medical image segmentation," *arXiv preprint arXiv:2503.01306*, 2025.

[132] Z. Wu, C. Shen, and A. v. d. Hengel, "Bridging category-level and instance-level semantic image segmentation," *arXiv preprint arXiv:1605.06885*, 2016.

[133] T.-Y. Lin *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[134] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[135] C. H. Sudre *et al.*, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248.

[136] K. C. Wong *et al.*, "3d segmentation with exponential logarithmic loss for highly unbalanced object sizes," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*. Springer, 2018, pp. 612–619.

[137] S. Shit *et al.*, "cldice - a novel topology-preserving loss function for tubular structure segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2021, p. 16555–16564. [Online]. Available: http://dx.doi.org/10.1109/CVPR46437.2021.01629

[138] J. Ma *et al.*, "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, p. 102035, 2021.

[139] B. Liu *et al.*, "Do we really need dice? the hidden region-size biases of segmentation losses," *Medical Image Analysis*, vol. 91, p. 103015, 2024.

[140] Z.-H. Feng *et al.*, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2235–2245.

[141] L. Maier-Hein *et al.*, "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nature communications*, vol. 9, no. 1, p. 5217, 2018.

[142] E. Christodoulou *et al.*, " Confidence intervals uncovered: Are we ready for real-world medical imaging AI? ," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. LNCS 15010. Springer Nature Switzerland, October 2024, pp. 124 – 132.

[143] K. Zou *et al.*, "A review of uncertainty estimation and its application in medical imaging," *Meta-Radiology*, vol. 1, no. 1, p. 100003, 2023.

[144] E. Kats, J. Goldberger, and H. Greenspan, "Soft labeling by distilling anatomical knowledge for improved ms lesion segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1563–1566.

[145] H. Li *et al.*, "Superpixel-guided label softening for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*. Springer, 2020, pp. 227–237.

[146] M. Gaillochet, C. Desrosiers, and H. Lombaert, "Active learning for medical image segmentation with stochastic batches," *Medical Image Analysis*, vol. 90, p. 102958, 2023.

[147] X. Li *et al.*, "Hal-ia: A hybrid active learning framework using interactive annotation for medical image segmentation," *Medical Image Analysis*, vol. 88, p. 102862, 2023.

[148] J. Ma *et al.*, "Medsam2: Segment anything in 3d medical images and videos," *arXiv preprint arXiv:2504.03600*, 2025.

[149] H. Guan and M. Liu, "Domain adaptation for medical image analysis: a survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2021.

[150] M. A. Morid, A. Borjali, and G. Del Fiol, "A scoping review of transfer learning research on medical image analysis using imagenet," *Computers in biology and medicine*, vol. 128, p. 104115, 2021.

[151] J. Connor *et al.*, "Reliability of sciseg automated measurement of midsagittal tissue bridges in spinal cord injuries using an external dataset," *Topics in Spinal Cord Injury Rehabilitation*, vol. 31, no. 2, pp. 39–49, 2025.

[152] S. Schading-Sassenhausen, V. Dietz, and P. Freund, "Effect of corticospinal and reticulospinal tract damage on spastic muscle tone and mobility: a retrospective observational mri study," *eBioMedicine*, vol. 118, p. 105824, 2025.

[153] E. Naga Karthik *et al.*, "Sciseg: Automatic segmentation of intramedullary lesions in spinal cord injury on t2-weighted mri scans," *Radiology: Artificial Intelligence*, vol. 7, no. 1, p. e240005, 2025, pMID: 39503603. [Online]. Available: https://doi.org/10.1148/ryai.240005

[154] G. Scivoletto *et al.*, "Acute traumatic and ischemic spinal cord injuries have a comparable course of recovery," *Neurorehabilitation and Neural Repair*, vol. 34, no. 8, pp. 723–732, Aug. 2020.

[155] P. W. New, R. A. Cripps, and B. Bonne Lee, "Global maps of non-traumatic spinal cord injury epidemiology: towards a living data repository," *Spinal Cord*, vol. 52, no. 2, pp. 97–109, Feb. 2014.

[156] E. Iseli *et al.*, "Prognosis and recovery in ischaemic and traumatic spinal cord injury: clinical and electrophysiological evaluation," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 67, no. 5, pp. 567–571, Nov. 1999.

[157] F. Miyanji *et al.*, "Acute cervical traumatic spinal cord injury: MR imaging findings correlated with neurologic outcome–prospective study with 100 consecutive patients," *Radiology*, vol. 243, no. 3, pp. 820–827, Jun. 2007.

[158] M. Dobran *et al.*, "Prognostic MRI parameters in acute traumatic cervical spinal cord injury," *European Spine Journal*, vol. 32, no. 5, pp. 1584–1590, May 2023.

[159] E. Huber *et al.*, "Are midsagittal tissue bridges predictive of outcome after cervical spinal cord injury?" *Annals of Neurology*, vol. 81, no. 5, pp. 740–748, May 2017.

[160] S. Kurpad *et al.*, "Impact of baseline magnetic resonance imaging on neurologic, functional, and safety outcomes in patients with acute traumatic spinal cord injury," *Global Spine Journal*, vol. 7, no. 3 Suppl, p. 151S, Sep. 2017.

[161] A. R. Martin *et al.*, "A novel MRI biomarker of spinal cord white matter injury: T2*-weighted white matter to gray matter signal intensity ratio," *American Journal of Neuroradiology*, vol. 38, no. 6, pp. 1266–1273, 2017.

[162] A. Bischof *et al.*, "Spinal cord atrophy predicts progressive disease in relapsing multiple sclerosis," *Annals of Neurology*, vol. 91, no. 2, pp. 268–281, Feb. 2022.

[163] N. Mummaneni *et al.*, "Injury volume extracted from MRI predicts neurologic outcome in acute spinal cord injury: A prospective TRACK-SCI pilot study," *Journal of Clinical Neuroscience*, vol. 82, no. Pt B, pp. 231–236, Dec. 2020.

[164] K. Vallotton *et al.*, "Width and neurophysiologic properties of tissue bridges predict recovery after cervical injury," *Neurology*, vol. 92, no. 24, pp. e2793–e2802, Jun. 2019.

[165] O. Khan *et al.*, "Predictive modeling of outcomes after traumatic and nontraumatic spinal cord injury using machine learning: Review of current progress and future directions," *Neurospine*, vol. 16, no. 4, pp. 678–685, Dec. 2019.

[166] N. Dietz *et al.*, "Machine learning in clinical diagnosis, prognostication, and management of acute traumatic spinal cord injury (SCI): A systematic review," *Journal of Clinical Orthopaedics and Trauma*, vol. 35, p. 102046, Dec. 2022.

[167] S. Asgari Taghanaki *et al.*, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137–178, Jan. 2021.

[168] D. B. McCoy *et al.*, "Convolutional neural network-based automated segmentation of the spinal cord and contusion injury: Deep learning biomarker correlates of motor impairment in acute spinal cord injury," *American Journal of Neuroradiology*, vol. 40, no. 4, pp. 737–744, Apr. 2019.

[169] R. Rupp *et al.*, "International standards for neurological classification of spinal cord injury: Revised 2019," *Topics in Spinal Cord Injury Rehabilitation*, vol. 27, no. 2, pp. 1–22, 2021.

[170] A. C. Smith *et al.*, "Spinal cord tissue bridges validation study: Predictive relationships with sensory scores following cervical spinal cord injury," *Topics in Spinal Cord Injury Rehabilitation*, vol. 28, no. 2, pp. 111–115, 2022.

[171] J. Cohen-Adad *et al.*, "Demyelination and degeneration in the injured human spinal cord detected with diffusion and magnetization transfer MRI," *Neuroimage*, vol. 55, no. 3, pp. 1024–1033, 2011.

[172] K. J. Gorgolewski *et al.*, "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments," *Scientific Data*, vol. 3, no. 1, p. 160044, Dec. 2016.

[173] F. Isensee *et al.*, "nnU-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[174] B. De Leener, S. Kadoury, and J. Cohen-Adad, "Robust, accurate and fast automatic segmentation of the spinal cord," *Neuroimage*, 2014.

[175] S. Bédard *et al.*, "Towards contrast-agnostic soft segmentation of the spinal cord," *arXiv [eess.IV]*, Oct. 2023.

[176] O. Commowick *et al.*, "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Scientific Reports*, vol. 8, no. 1, p. 13650, Sep. 2018.

[177] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020.

[178] C. Blanc *et al.*, "Combining PropSeg and a convolutional neural network for automatic spinal cord segmentation in pediatric populations and patients with spinal cord injury," *International Journal of Imaging Systems and Technology*, vol. n/a, no. n/a, Feb. 2023.

[179] A. Nouri *et al.*, "The relationship between MRI signal intensity changes, clinical presentation, and surgical outcome in degenerative cervical myelopathy," *Spine*, vol. 42, no. 24, pp. 1851–1858, 2017.

[180] A. R. Martin *et al.*, "Imaging evaluation of degenerative cervical myelopathy: Current state of the art and future directions," pp. 33–45, Jan. 2018.

[181] P. S. Scheuren *et al.*, "Combined neurophysiologic and neuroimaging approach to reveal the structure-function paradox in cervical myelopathy," *Neurology*, Aug. 2021.

[182] A. C. Smith *et al.*, "Ambulatory function in motor incomplete spinal cord injury: a magnetic resonance imaging study of spinal cord edema and lower extremity muscle morphometry," *Spinal Cord*, vol. 55, no. 7, pp. 672–678, Jul. 2017.

[183] D. P. Cummins *et al.*, "Establishing the inter-rater reliability of spinal cord damage manual measurement using magnetic resonance imaging," *Spinal Cord Series and Cases*, vol. 5, p. 20, Feb. 2019.

[184] E. N. Karthik *et al.*, "SCIsegV2: A universal tool for segmentation of intramedullary lesions in spinal cord injury," *arXiv [cs.CV]*, Jul. 2024.

[185] F. Barkhof *et al.*, "Comparison of MRI criteria at first presentation to predict conversion to clinically definite multiple sclerosis," *Brain*, vol. 120 (Pt 11), pp. 2059–2069, Nov. 1997.

[186] C. Trolle, E. Goldberg, and C. Linnman, "Spinal cord atrophy after spinal cord injury – a systematic review and meta-analysis," *NeuroImage: Clinical*, vol. 38, p. 103372, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221315822300061X

[187] A. R. Martin *et al.*, "Monitoring for myelopathic progression with multiparametric quantitative MRI," *PLoS One*, vol. 13, no. 4, p. e0195733, Apr. 2018.

[188] B. De Leener *et al.*, "Segmentation of the human spinal cord," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 29, no. 2, pp. 125–153, Apr. 2016.

[189] M. A. Horsfield *et al.*, "Rapid semi-automatic segmentation of the spinal cord from magnetic resonance images: application in multiple sclerosis," *Neuroimage*, vol. 50, no. 2, pp. 446–455, Apr. 2010.

[190] G. Kim *et al.*, "T1- vs. t2-based MRI measures of spinal cord volume in healthy subjects and patients with multiple sclerosis," *BMC Neurology*, vol. 15, no. 1, p. 124, Jul. 2015.

[191] B. Billot *et al.*, "A learning strategy for contrast-agnostic MRI segmentation," Mar. 2020.

[192] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, Mar. 2022.

[193] K. Kamnitsas *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25*. Springer, 2017, pp. 597–609.

[194] Q. Dou *et al.*, "Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation," *IEEE Access*, vol. 7, pp. 99 065–99 076, 2019.

[195] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[196] J. Hoffman *et al.*, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. Pmlr, 2018, pp. 1989–1998.

[197] C. Chen *et al.*, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 865–872.

[198] H. Li *et al.*, "Domain generalization for medical imaging classification with linear-dependency regularization," *ArXiv*, vol. abs/2009.12829, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221970126

[199] R. Zhang *et al.*, "Semi-supervised domain generalization for medical image analysis," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–5.

[200] H. Oliveira *et al.*, "Domain generalization in medical image segmentation via meta-learners," in *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, vol. 1, 2022, pp. 288–293.

[201] Z. Zhang *et al.*, "Domain generalization with adversarial intensity attack for medical image segmentation," *arXiv preprint arXiv:2304.02720*, 2023.

[202] M. Bateson *et al.*, "Source-free domain adaptation for image segmentation," *Medical Image Analysis*, vol. 82, p. 102617, Nov. 2022.

[203] A. Zhao *et al.*, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8543–8553.

[204] C. Ouyang *et al.*, "Causality-inspired single-source domain generalization for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 4, pp. 1095–1106, 2023.

[205] H. Chaves *et al.*, "Brain volumes quantification from MRI in healthy controls: Assessing correlation, agreement and robustness of a convolutional neural network-based software against FreeSurfer, CAT12 and FSL," *Journal of Neuroradiology*, vol. 48, no. 3, pp. 147–156, May 2021.

[206] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances of Neural Information Processing Systems*, vol. 32, 2019.

[207] M. Jorge Cardoso *et al.*, "MONAI: An open-source framework for deep learning in healthcare," Nov. 2022.

[208] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," Jul. 2016.

[209] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," http://robotics.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf, 2013, accessed: 2023-9-18.

[210] Q. Dou *et al.*, "3D deeply supervised network for automated segmentation of volumetric medical images," *Medical Image Analysis*, vol. 41, pp. 40–54, Oct. 2017.

[211] R. Deng *et al.*, "Learning to predict crisp boundaries," in *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 570–586.

[212] J. Bertels *et al.*, "Optimization with soft dice can lead to a volumetric bias," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, 2020, pp. 89–97.

[213] S. Jia *et al.*, "Automatically segmenting the left atrium from cardiac images using successive 3D U-Nets and a contour loss," in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges*. Springer International Publishing, 2019, pp. 221–229.

[214] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *International workshop on machine learning in medical imaging*. Springer, 2017, pp. 379–387.

[215] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6971–6981.

[216] C. Kaul *et al.*, "Penalizing small errors using an adaptive logarithmic loss," in *Pattern Recognition. ICPR International Workshops and Challenges*. Springer International Publishing, 2021, pp. 368–375.

[217] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014.

[218] B. Benjdira *et al.*, "Spinal cord segmentation in ultrasound medical imagery," *MDPI: Applied Sciences*, vol. 10, no. 4, p. 1370, Feb. 2020.

[219] A. Lemay *et al.*, "Automatic multiclass intramedullary spinal cord tumor segmentation on mri with deep learning," *NeuroImage: Clinical*, vol. 31, p. 102766, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213158221002102

[220] A. Kirillov *et al.*, "Segment anything," *arXiv:2304.02643*, 2023.

[221] E. N. Karthik *et al.*, "Contrast-agnostic spinal cord segmentation: A comparative study of convnets and vision transformers," in *Medical Imaging with Deep Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=n6D25aqdV3

[222] ——, "Segmentation of multiple sclerosis lesions across hospitals: Learn continually or train from scratch?" *arXiv [cs.CV]*, Oct. 2022.

[223] N. A. Losseff *et al.*, "Spinal cord atrophy and disability in multiple sclerosis. a new reproducible and sensitive MRI method with potential to monitor disease progression," *Brain*, vol. 119 (Pt 3), pp. 701–708, Jun. 1996.

[224] C. Lukas *et al.*, "Relevance of spinal cord abnormalities to clinical disability in multiple sclerosis: MR imaging findings in a large cohort of patients," *Radiology*, vol. 269, no. 2, pp. 542–552, Nov. 2013.

[225] M. Horáková *et al.*, "Semi-automated detection of cervical spinal cord compression with the spinal cord toolbox," *Quantitative Imaging in Medicine and Surgery*, vol. 12, no. 4, pp. 2261–2279, Apr. 2022.

[226] J. Valošek and J. Cohen-Adad, "Reproducible spinal cord quantitative mri analysis with the spinal cord toolbox," *Magnetic Resonance in Medical Sciences*, vol. 23, no. 3, pp. 307–315, 2024.

[227] M. Chen *et al.*, "Automatic magnetic resonance spinal cord segmentation with topology constraints for variable fields of view," *NeuroImage*, vol. 83, pp. 1051–1062, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S105381191300829X

[228] N. Masse-Gignac *et al.*, "Attention-gated U-net networks for simultaneous axial/sagittal planes segmentation of injured spinal cords," *Journal of Applied Clinical Medical Physics*, vol. 24, no. 10, p. e14123, Oct. 2023.

[229] C. Tsagkas *et al.*, "Fully automatic method for reliable spinal cord compartment segmentation in multiple sclerosis," *American Journal of Neuroradiology*, vol. 44, no. 2, pp. 218–227, Feb. 2023.

[230] L.-C. Chen *et al.*, "Rethinking atrous convolution for semantic image segmentation," *arXiv [cs.CV]*, Jun. 2017.

[231] M. Taso *et al.*, "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*, vol. 29, no. 6, pp. 817–832, Jun. 2016.

[232] N. Papinutto *et al.*, "Intersubject variability and normalization strategies for spinal cord total cross-sectional and gray matter areas," *Journal of Neuroimaging*, vol. 30, no. 1, pp. 110–118, Jan. 2020.

[233] S. Bédard and J. Cohen-Adad, "Automatic measure and normalization of spinal cord cross-sectional area using the pontomedullary junction," *Frontiers in Neuroimaging*, vol. 1, p. 1031253, Nov. 2022.

[234] R. Labounek *et al.*, "Body size interacts with the structure of the central nervous system: A multi-center in vivo neuroimaging study," *bioRxiv*, May 2024.

[235] F. Kato *et al.*, "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*, 2012.

[236] I. Prapas *et al.*, "Continuous training and deployment of deep learning models," *Datenbank Spektrum*, vol. 21, no. 3, pp. 203–212, Nov. 2021.

[237] E. Agirre, A. Jonsson, and A. Larcher, "Framing lifelong learning as autonomous deployment: Tune once live forever," in *Lecture Notes in Electrical Engineering*, ser. Lecture notes in electrical engineering. Singapore: Springer Singapore, 2021, vol. 0, pp. 331–336.

[238] B. Liu and S. Mazumder, "Lifelong and continual learning dialogue systems: Learning during conversation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 15 058–15 063, May 2021.

[239] M. Treveil *et al.*, *Introducing MLOps*. Sebastopol, CA: O'Reilly Media, Dec. 2020.

[240] S. Alla and S. K. Adari, *Beginning MLOps with MLFlow: Deploy models in AWS SageMaker, Google cloud, and Microsoft azure*, 1st ed. Berlin, Germany: APress, Dec. 2020.

[241] A. I. U. Tabassam, "MLOps: A step forward to enterprise machine learning," *arXiv [cs.SE]*, May 2023.

[242] N. Molinier and S. Bédard, "whole-spine," 2024.

[243] J. Shi and J. Wu, "Distilling effective supervision for robust medical image segmentation with noisy labels," *arXiv [cs.CV]*, Jun. 2021.

[244] J. Yao *et al.*, "Learning to segment from noisy annotations: A spatial correction approach," *arXiv [eess.IV]*, Jul. 2023.

[245] A. Rahman *et al.*, "Power mean based image segmentation in the presence of noise," *Scientific Reports*, vol. 12, no. 1, p. 21177, Dec. 2022.

[246] A. S. Jwa, M. Norgaard, and R. A. Poldrack, "Can I have your data? recommendations and practical tips for sharing neuroimaging data upon a direct personal request," *Imaging Neuroscience*, vol. 3, Mar. 2025.

[247] E. Naga Karthik *et al.*, "Automatic segmentation of spinal cord lesions in ms: A robust tool for axial t2-weighted mri scans," *Imaging Neuroscience*, vol. 3, pp. IMAG–a, 2025.

[248] O. Spjuth, J. Frid, and A. Hellander, "The machine learning life cycle and the cloud: implications for drug discovery," *Expert Opinion on Drug Discovery*, vol. 16, no. 9, pp. 1071–1079, Sep. 2021.

[249] N. Kandpal *et al.*, "Git-theta: A git extension for collaborative development of machine learning models," *arXiv [cs.LG]*, Jun. 2023.

[250] C. González *et al.*, "Regulating radiology AI medical devices that evolve in their life-cycle," *arXiv [cs.CY]*, Dec. 2024.

[251] B. Joo *et al.*, "Establishing normative values for entire spinal cord morphometrics in east asian young adults," *Korean Journal of Radiology*, vol. 26, no. 2, pp. 146–155, Feb. 2025.

[252] I. Shumailov *et al.*, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, no. 8022, pp. 755–759, Jul. 2024.

[253] J. J. M. van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, Nov. 2017.

[254] L. Yang *et al.*, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. Springer, 2017, pp. 399–407.

[255] J. Oh *et al.*, "The canadian prospective cohort study to understand progression in multiple sclerosis (canproco): rationale, aims, and study design," *BMC neurology*, vol. 21, pp. 1–19, 2021.

[256] M. Antonelli *et al.*, "The medical segmentation decathlon," *Nature communications*, vol. 13, no. 1, p. 4128, 2022.

[257] A. Hatamizadeh *et al.*, "Unetr: Transformers for 3d medical image segmentation," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1748–1758, 2021.

[258] Z. Ke *et al.*, "Continual pre-training of language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.03241

[259] Çağatay Yıldız *et al.*, "Investigating continual pretraining in large language models: Insights and implications," 2025. [Online]. Available: https://arxiv.org/abs/2402.17400

[260] V. Oliva *et al.*, "Mapping hand function with simultaneous brain-spinal cord functional mri," *bioRxiv*, 2025. [Online]. Available: https://www.biorxiv.org/content/early/2025/02/27/2025.02.27.640504

[261] J. Sohl-Dickstein *et al.*, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265.

[262] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[263] W. Peebles *et al.*, "Learning to learn with generative models of neural network checkpoints," *arXiv preprint arXiv:2209.12892*, 2022.

[264] A. Kline *et al.*, "Multimodal machine learning in precision health: A scoping review," *npj Digital Medicine*, vol. 5, no. 1, p. 171, 2022.

[265] D. Schouten *et al.*, "Navigating the landscape of multimodal ai in medicine: a scoping review on technical challenges and clinical applications," *arXiv preprint arXiv:2411.03782*, 2024.

[266] C. Li *et al.*, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *Advances in Neural Information Processing Systems*, vol. 36, pp. 28 541–28 564, 2023.

[267] D. Duenias *et al.*, "Hyperfusion: A hypernetwork approach to multimodal integration of tabular and medical imaging data for predictive modeling," *Medical Image Analysis*, vol. 102, p. 103503, May 2025. [Online]. Available: http://dx.doi.org/10.1016/j.media.2025.103503

[268] A. A. Barr, R. Rozman, and E. Guo, "Generative adversarial networks vs large language models: a comparative study on synthetic tabular data generation," *arXiv preprint arXiv:2502.14523*, 2025.

[269] J. Valošek *et al.*, "Automatic segmentation of the spinal cord nerve rootlets," *Imaging Neuroscience*, vol. 2, pp. 1–14, 07 2024. [Online]. Available: https://doi.org/10.1162/imag_a_00218

[270] Y. Warszawer *et al.*, "Totalspineseg: Robust spine segmentation with landmark-based labeling in mri," 03 2025.

[271] A. Hoopes *et al.*, "Voxelprompt: A vision-language agent for grounded medical image analysis," 2024. [Online]. Available: https://arxiv.org/abs/2410.08397

[272] A. Fallahpour *et al.*, "Medrax: Medical reasoning agent for chest x-ray," 2025. [Online]. Available: https://arxiv.org/abs/2502.02673

[273] M. Seif *et al.*, "Cervical cord neurodegeneration in traumatic and Non-Traumatic spinal cord injury," *Journal of Neurotrauma*, vol. 37, no. 6, pp. 860–867, Mar. 2020.

[274] D. R. O'Dell *et al.*, "Midsagittal tissue bridges are associated with walking ability in incomplete spinal cord injury: A magnetic resonance imaging case series," *Journal of Spinal Cord Medicine*, vol. 43, no. 2, pp. 268–271, Mar. 2020.

[275] D. Pfyffer *et al.*, "Prognostic value of tissue bridges in cervical spinal cord injury: a longitudinal, multicentre, retrospective cohort study," *The Lancet Neurology*, 2024.

[276] C. Matsoukas *et al.*, "Is it time to replace cnns with transformers for medical images?" *ArXiv*, vol. abs/2108.09038, 2021.

[277] L. Deininger *et al.*, "A comparative study between vision transformers and cnns in digital pathology," *ArXiv*, vol. abs/2206.00389, 2022.

[278] A. Fanizzi *et al.*, "Comparison between vision transformers and convolutional neural networks to predict non-small lung cancer recurrence," *Scientific Reports*, vol. 13, no. 1, p. 20605, 2023.

[279] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6971–6981.

[280] S. Roy *et al.*, "Mednext: Transformer-driven scaling of convnets for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023.

[281] A. Hatamizadeh *et al.*, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," *ArXiv*, vol. abs/2201.01266, 2022.

[282] C. Zeng *et al.*, "Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain mri," *Frontiers in Neuroinformatics*, vol. 14, 2020.

[283] F. La Rosa *et al.*, "Multiple sclerosis cortical and wm lesion segmentation at 3t mri: a deep learning method based on flair and mp2rage," *NeuroImage: Clinical*, vol. 27, p. 102335, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213158220301728

[284] C. Baweja, B. Glocker, and K. Kamnitsas, "Towards continual learning in medical imaging," *ArXiv*, vol. abs/1811.02496, 2018.

[285] K. A. van Garderen *et al.*, "Towards continuous learning for glioma segmentation with elastic weight consolidation," *ArXiv*, vol. abs/1909.11479, 2019.

[286] J. Hofmanninger *et al.*, "Dynamic memory to alleviate catastrophic forgetting in continuous learning settings," *ArXiv*, vol. abs/2007.02639, 2020.

[287] G. M. van de Ven and A. S. Tolias, "Three scenarios for continual learning," *ArXiv*, vol. abs/1904.07734, 2019.

[288] A. Kerbrat *et al.*, "Multiple sclerosis lesions in motor tracts from brain to cervical cord: spatial distribution and correlation with disability," *Brain*, vol. 143, no. 7, pp. 2089–2105, 06 2020. [Online]. Available: https://doi.org/10.1093/brain/awaa162

[289] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *NIPS*, 2017.

APPENDIX A

Automatic measurement of tissue bridges in SCI

Traumatic and non-traumatic SCI commonly involve intramedullary lesions, which are critical areas of tissue damage within the SC. MRI is routinely used to provide information on the extent and the location of these intramedullary lesions [34, 121, 273]. Importantly, MRI scans can also be used to compute quantitative biomarkers, such as midsagittal tissue bridges [159]. These help in quantifying the amount of preserved SC neural tissue (carrying motor and sensory information to and from the brain) and have been found to predict functional recovery in patients with traumatic and non-traumatic SCI [101, 109, 159, 170, 274, 275].

Quantifying tissue bridges

The manual measurement of tissue bridges is performed on a single midsagittal slice of a volumetric (3D) T2w MRI image [101, 109, 159, 164, 170, 274, 275] (Figure A.1A). The midsagittal slice is defined as the middle slice of all slices where the SC is visible (Figure A.1B). Ventral and dorsal tissue bridges are quantified as the width of spared tissue at the minimum distance from the intramedullary lesion edge to the boundary between the SC and cerebrospinal fluid (Figure A.1C).

To automate the measurement of tissue bridges, we propose a method that computes ventral and dorsal tissue bridges utilizing the lesion and SC segmentation masks. To compensate for different neck positions and, consequently, different SC curvatures, we use angle correction, which adjusts the tissue bridge widths with respect to the SC centerline. The method computes tissue bridges from all sagittal slices containing the lesion, allowing quantification of not only midsagittal but parasagittal tissue bridges as well. For the purpose of this study (and to compare against existing manual measurements based on midsagittal tissue bridges), we considered only the midsagittal slice for the automatic measurement of the tissue bridges.

Evaluation To validate the automatic measurements of the tissue bridges, we compared the method against manual and semi-automatic techniques in 15 individuals with traumatic SCI from site 1. Specifically, we compared the following: (1) **manual** — manual measurement of tissue bridges on manually segmented intramedullary lesions, (2) **semi-automatic** — automatic measurement of the tissue bridges using the proposed method on manually segmented intramedullary lesions, and (3) **fully-automatic** — automatic measurement of

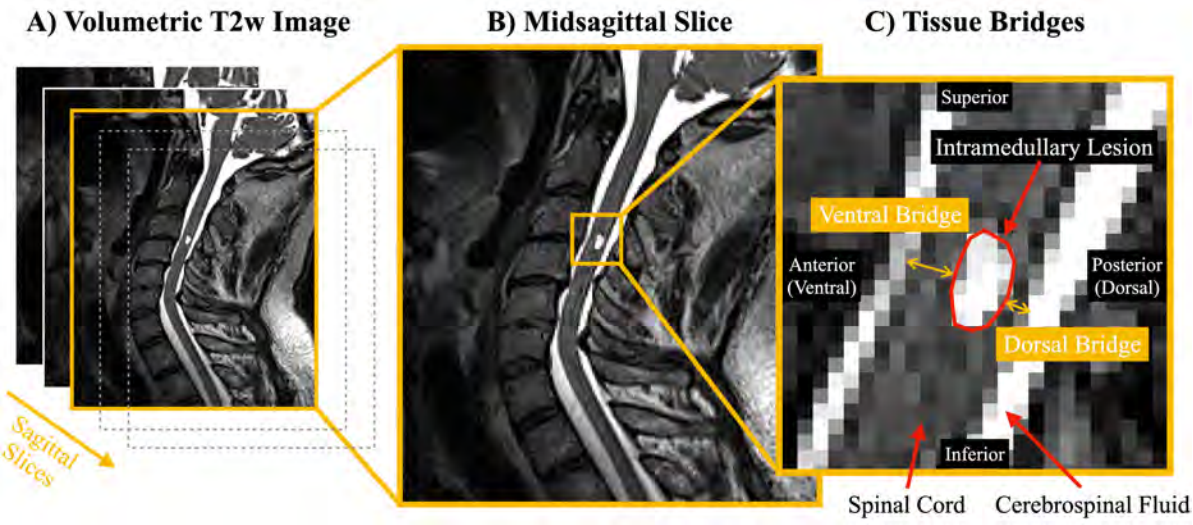

Figure A.1 Illustration of tissue bridges. A) Volumetric T2w image of a spinal cord injury (SCI) with chronic intramedullary lesion. B) Midsagittal slice used to compute the tissue bridges. C) Ventral and dorsal tissue bridges are defined as the width of spared tissue at the minimum distance from the intramedullary lesion edge to the boundary between the SC and cerebrospinal fluid.

tissue bridges using `SCIsegV2` predictions. Statistical analysis was performed using the SciPy v1.10.0. The distribution of the data was assessed with the D'Agostino and Pearson normality test. Subsequently, the Kruskal-Wallis H-test was performed to compare the methods independently for ventral and dorsal bridges.

Results Table A.1 shows the comparison of the midsagittal tissue bridges obtained using different methods (manual vs semi-automatic vs fully-automatic) for 15 patients with traumatic SCI from site 1. For the fully-automatic technique, we used the `SCIseg` model to obtain the lesion segmentations. There was *no* statistically significant ($p > .05$) difference between the bridges computed using different methods.

Discussion Automatic segmentation of the lesions could mitigate the bottleneck and inter-rater variability associated with manual annotations. Likewise, automatically measuring tissue bridges could provide an objective, unbiased way in guiding rehabilitation decision making and stratifying patients into homogeneous subgroups of recovery in clinical trials. While the proposed proof-of-concept only measures the midsagittal tissue bridges (as it is the current standard operating procedure), a comprehensive evaluation of the width of the spared tissue bridges can be obtained by combining measurements from both *parasagittal* and *midsagittal* slices.

Table A.1 Comparison of *ventral* and *dorsal* midsagittal tissue bridges between manual, semi-automatic, and automatic measurements. Values are reported in millimetres.

ID	Manual Lesions & Manual Measurements		Manual Lesions & Automatic Measurements		`SCIseg` Predictions & Automatic Measurements	
	Ventral	**Dorsal**	**Ventral**	**Dorsal**	**Ventral**	**Dorsal**
sub-zh101	0	2.65	0	2.39	0.34	2.39
sub-zh102	2.10	0.83	2.25	0	2.27	0.67
sub-zh104	0	0	0.54	0	0.55	0
sub-zh105	2.70	0	2.38	0.60	2.99	0
sub-zh106	0	0	0	0	0	0
sub-zh107	0	0.76	0	0.67	0	0.65
sub-zh108	1.32	0.52	1.96	0.66	2.03	0.68
sub-zh109	1.13	1.03	0.71	0	1.08	0.73
sub-zh110	0	0.99	0	0	0	0.39
sub-zh112	3.01	0.36	1.70	0.44	2.64	0.44
sub-zh114	0	0	0	0.38	0	0
sub-zh115	0	0	0	2.12	0	0.42
sub-zh116	3.12	0.50	2.38	0	2.49	0.80
sub-zh118	0.40	0	0	0	0	0
sub-zh119	2.93	2.98	1.04	0.50	1.48	0.95

APPENDIX B

Supplementary material: Towards contrast-agnostic soft segmentation of the spinal cord

Pairwise correlation plots for all six MR contrasts

Figure B.1 shows the r^2 correlation plot between CSA at C2-C3 vertebral levels for each pair of contrasts used for the proposed `soft_all` model. We observe a high level of agreement between MT-on / DWI (row 1, column 1), GRE-T1w / MT-on contrasts (row 2, column 2), T2*w / MT-on (row 4, column 2), T2*w / GRE-T1w (row 4, column 3), and T2w / T1w contrasts (row 5, column 4).

Baselines: Comparison between absolute CSA error per contrast

Figures B.2 to B.4 show the absolute CSA error across MRI contrasts for each baseline, except for the error plot of the proposed `soft_all` model which is presented in Figure 5.2 in Chapter 5.

Soft vs. Hard ground-truth masks

Comparing `hard_all_SoftSeg` (Figure B.2) and `soft_all` (Figure 5.2), the difference between the two methods can be explained by the fact that `hard_all_SoftSeg` was trained with binary (hard) GT masks, whereas `soft_all` was trained with averaged soft GT masks, both using the adaptive wing loss. We notice larger absolute CSA errors for the `hard_all_SoftSeg` model, especially for the T2*w, DWI and MT-on contrasts. Considering that these 3 contrasts are acquired with thick axial slices (and hence suffer from higher partial volume compared to the T1w and T2w), it is likely that the discrepancy between these models highlights the capability of our proposed procedure for the creation of soft GT masks for training that better encode partial volume information. Note that the DWI contrast is an average across diffusion directions after motion correction. Despite the fact that motion correction is applied, slight residual motion across volumes blurs the edges of the spinal cord when averaging the volumes, resulting in higher partial volume. This could also explain the higher CSA error of `hard_all_SoftSeg` compared to `soft_all` for DWI contrasts.

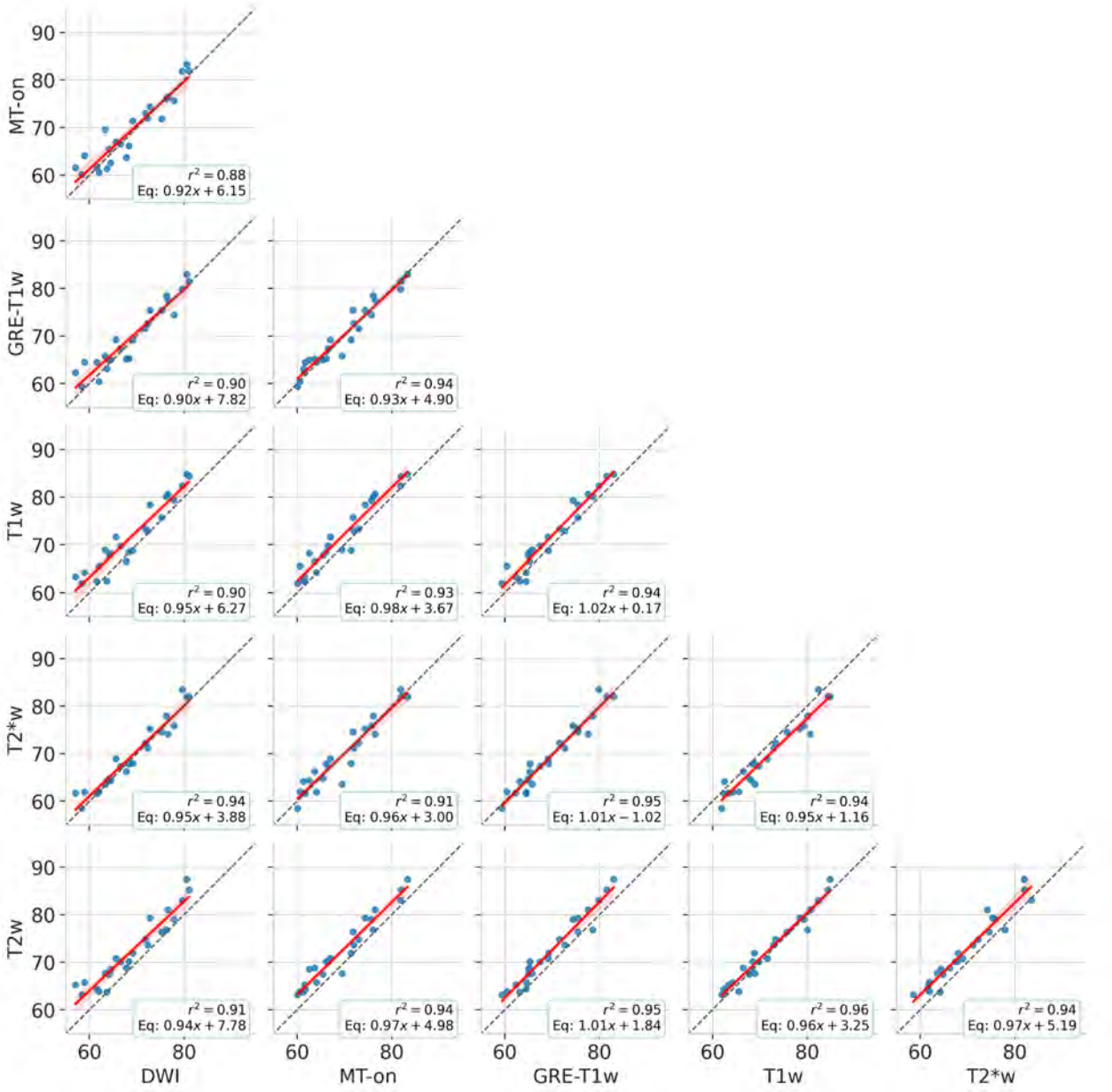

Figure B.1 Pairwise correlation plots showing the level of agreement between CSA for each pair of contrasts for the proposed `soft_all` model. Each scatter point represents one participant and the dashed line corresponds perfect agreement.

Dice cross-entropy vs. adaptive wing loss

As mentioned in Section 2.3.4 of the main manuscript, using Dice coefficient in the training objective does not optimize for the accuracy of the segmentations at the spinal cord/cerebrospinal fluid boundary. This led to subtle under-segmentations across participants, thereby resulting in consistently larger absolute CSA errors by more than an order of 1 mm^2 across

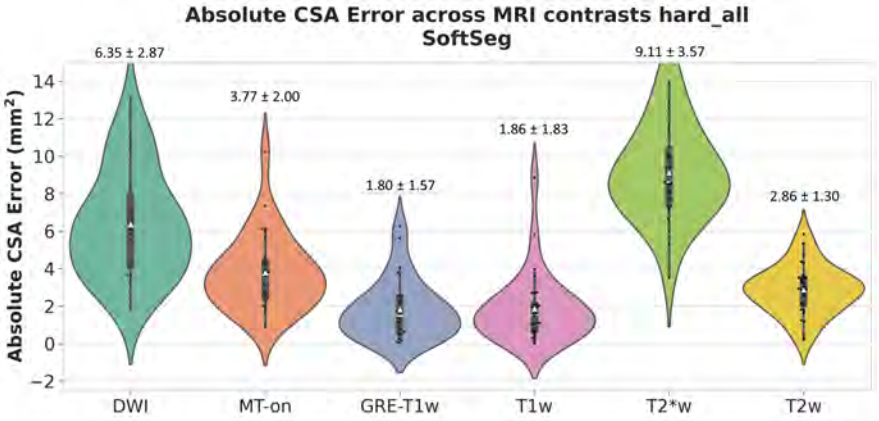

Figure B.2 Absolute CSA error between the predictions and GT across each contrast for the `hard_all_SoftSeg` model trained on all contrasts with hard GT masks. Scatter plots within each violin represent the individual CSA errors for all test participants. White triangle marker shows the mean CSA error.

all contrasts. Figures B.3 and B.4 show the CSA errors per contrasts for models trained with the DiceCE loss using hard masks and averaged soft masks, respectively. As expected, for the `hard_all_diceCE_loss` model, the individual CSA estimations across contrasts vary substantially. For the `soft_all_diceCE_loss` model, we see a relative improvement across contrasts, but does not outperform the `soft_all` model shown in Figure 5.2 in Chapter 5.

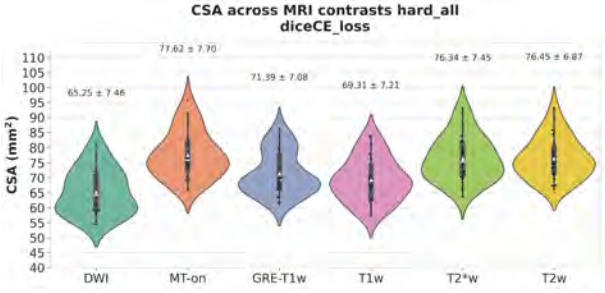

Figure B.3 Absolute CSA error between the predictions and GT across each contrast for the model trained on all contrasts with hard GT masks and Dice cross-entropy loss (instead of adaptive wing loss). Scatter plots within each violin represent the individual CSA errors for all test participants. White triangle marker shows the mean CSA error.

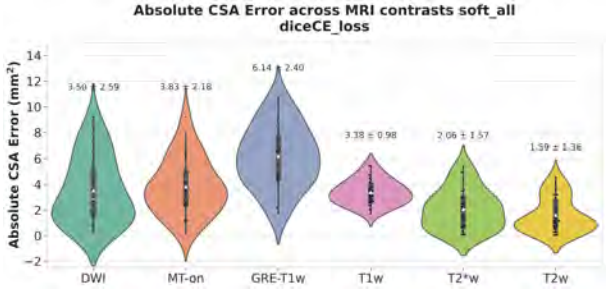

Figure B.4 Absolute CSA error between the predictions and GT across each contrast for the model trained on all contrasts with soft GT masks and Dice cross-entropy loss (instead of adaptive wing loss). Scatter plots within each violin represent the individual CSA errors for all test participants. White triangle marker shows the mean CSA error.

State of the art: CSA variability across all methods

Figures B.5 and B.6 show the variability of CSA across *all* methods. Note that subsets of these plots are reported in Chapter 5. Considering a comparison within the DeepSeg models, we see that the 2D model achieves relatively better results (i.e. lower CSA errors) than the 3D model. The worse performance of the 3D model can be explained by the *patch size* chosen for inference with sliding windows. In the source code, we observed that the patch sizes were fixed to $64 \times 64 \times 48$ and $96 \times 96 \times 48$ depending on the contrast, which do not contain enough contextual information of the cord in the A-P and S-I axes. This means that these patch sizes unintentionally cut off patches of the cord, thereby not providing its complete structure. On the other hand, the 2D model uses individual slices in the S-I axes containing the complete cross-sectional view of the cord during inference. This results in a superior performance of the DeepSeg 2D compared to DeepSeg 3D.

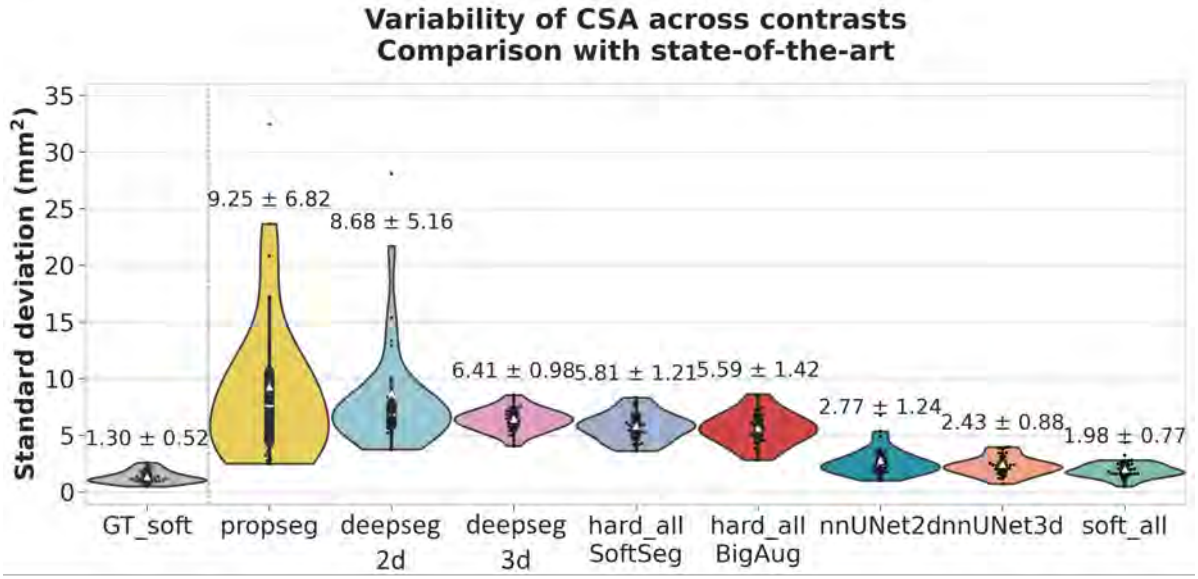

Figure B.5 Standard deviation of CSA between C2-C3 vertebral levels for PropSeg, DeepSeg3D/2D, `hard_all_SoftSeg`, `hard_all_BigAug`, nnUNet3D/2D, and our model `soft_all`. White triangle marker shows the mean CSA STD.

With the nnUNet models, nnUNet2D model performed slightly better in terms of the absolute CSA error (i.e. the error was slightly lower) and slightly worse standard deviation (STD) of cross-sectional area (CSA) across contrasts, compared to nnUNet3D. Within the nnUNet models, nnUNet3D used a patch size of 80 × 192 × 160 (RPI orientation) for training, while nnUNet2D used a patch size of 256 × 224 (PI orientation) slicing up the 3D volume along the R-L dimension. A larger patch-size, especially in the superior-inferior dimension

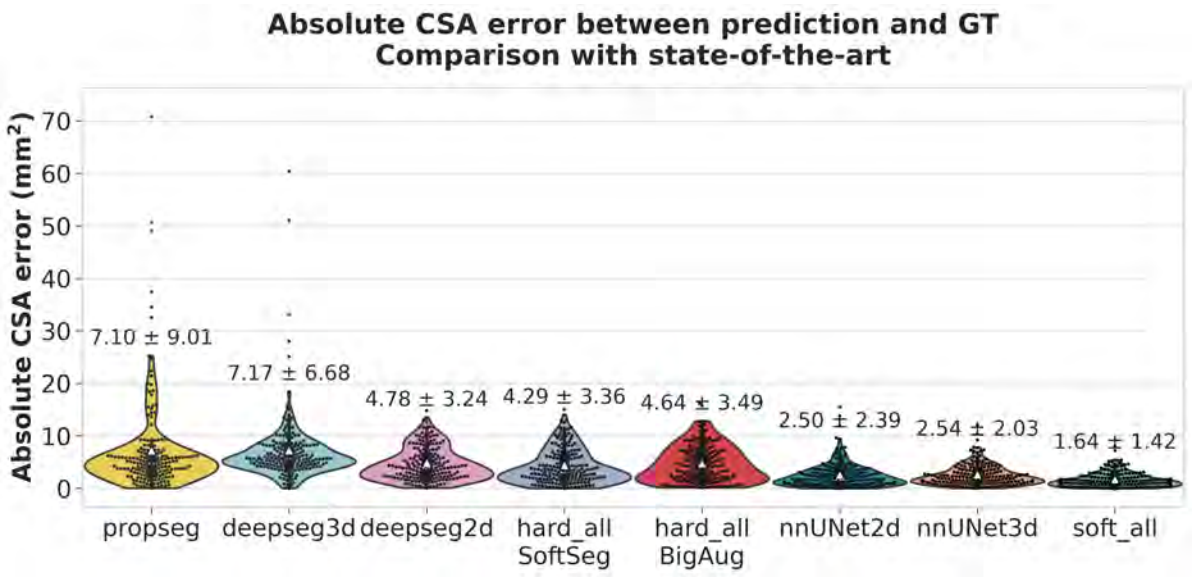

Figure B.6 Mean absolute CSA error for PropSeg, DeepSeg3D/2D, `hard_all_SoftSeg`, `hard_all_BigAug`, nnUNet3D/2D, and our model `soft_all`. White triangle marker showsthe mean CSA error.

(160 in 3D vs. 224 in 2D) resulted in the 2D model performing slightly better than the 3D model. Furthermore, the difference between nnUNet 2D and 3D is not as substantial as the one observed DeepSeg 2D and 3D because DeepSeg3D used a much smaller patch size ($64 \times 64 \times 48$ or $96 \times 96 \times 48$), failing to treat the (tubular) spinal cord structure as a whole.

Lastly, nnUNet2D performed considerably better than DeepSeg2D mainly because it was trained on soft masks that were binarized at 0.5 threshold, further emphasizing that our proposed preprocessing pipeline for creating soft masks by combining multiple contrasts is key to reducing morphometric variability across contrasts.

Contrast-agnostic segmentation: Ablations across contrasts

In this section, we performed two ablation studies to evaluate the stability of the model with respect to the number of contrasts used to train the model. For each ablation study, we preprocessed the data to generate a unique soft segmentation GT that only included the selected contrasts. The models were then trained using the same parameters as our `soft_all` model for both ablations.

For the first ablation study, we trained the model with two contrasts only: T1w and T2w. These were chosen because of their large field-of-view (FOV) compared to the remaining contrasts. Panel A of Figures B.7 and B.8 respectively show the average CSA and CSA error

across all contrasts from the test set. As the model was trained on the soft GT averaged from T1w and T2w, we expected to see little divergence in CSA for these two contrasts in the test set, which is indeed confirmed. Interestingly, the model also performed reasonably well on the MTon and GRE-T1w contrasts in terms of CSA estimate, which is likely due to the similar cord/CSF appearance between MTon and T2w and between GRE-T1w and T1w. That being said, the absolute mean error for MT-on (7.01 mm^2) and GRE-T1w (7.82 mm^2) is much larger than that for T1w (1.33 mm^2) and T2w (1.28 mm^2). However, the model clearly does not perform well for DWI and T2*w contrasts.

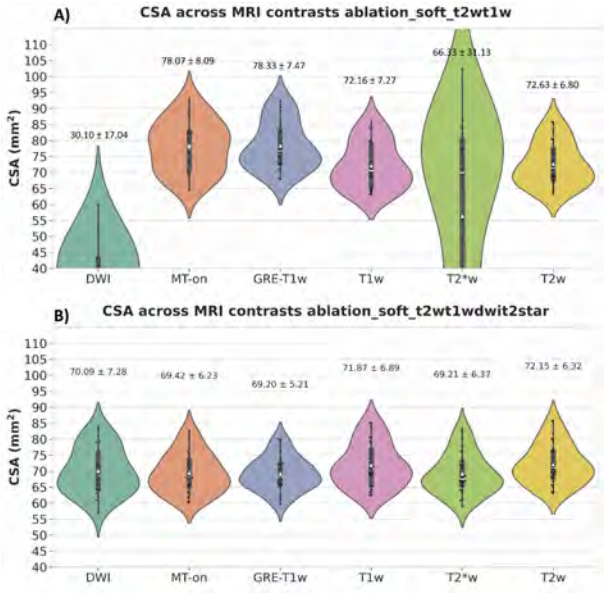

Figure B.7 Effect of number of contrasts included in the GT and training on CSA. A) CSA values of test set for a model that include T1w and T2w contrasts. B) CSA values of test set for a model that include T1w, T2w, DWI and T2*w contrasts. White triangle marker shows the mean CSA across participants.

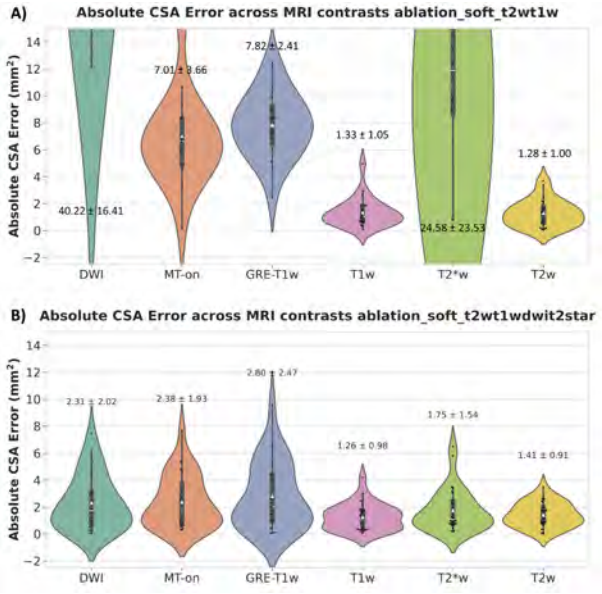

Figure B.8 Absolute CSA error between the predictions and GT for the `soft_all` model including T1w and T2w contrasts (A) and for the `soft_all` model including T1w, T2w, DWI and T2*w contrasts (B). Scatter plots within each violin represent the individual CSA errors for all participants in the test set. White triangle marker shows the mean CSA error across participants.

For the second ablation study, we trained the model with four contrasts: T1w, T2w, DWI and T2*w. There, we expected results to be more favourable to the DWI and T2*w contrasts, which is indeed confirmed by Figures B.7 and B.8, panel B. Overall, we observe that CSA values across contrasts (Figure B.7B) are more similar across all 6 contrasts, even if only 4 were used for creating the GT masks and model training. It is important to note that the number of images in the training set doubled compared to the first ablation (T1w and T2w).

SynthSeg for spinal cord segmentation

SynthSeg was originally proposed for the segmentation of brain scans of any resolution and contrast, however, it requires fully-labeled scans to synthesize brain images by sampling from a Gaussian Mixture Model (GMM). These synthetic images are then used to train the segmentation model. As described in Section 5.4 of the original paper showing an extension of SynthSeg to cardiac segmentation, the labels for all tissues (i.e. of those regions other than the anatomy of interest) were obtained by clustering the intensities in the image using the Expectation Maximization (EM) algorithm.

As SynthSeg used labels from T1w scans for generating the synthetic scans (see Table 1), we also used T1w isotropic spinal cord MRI scans for a fair comparison. Starting with GT spinal cord segmentation masks, we used the K-Means clustering algorithm to generate labels for additional structures (i.e. everything outside the spinal cord) ranging between 3-10 clusters. Specifically, keeping the background and the spinal cord labels fixed (0 and 1, respectively), automated labels for the rest of the T1w scan were obtained by clustering the corresponding intensities into N classes ($N \in [3, 10]$, as done in the paper). During training, one of these enhanced label maps is randomly selected to synthesize a training image. In total, 1552 automatically obtained labels, from T1w scans of 194 subjects, were used for training. Figure B.9 shows the manual GT along with the automatic K-Means labels and a sample training image.

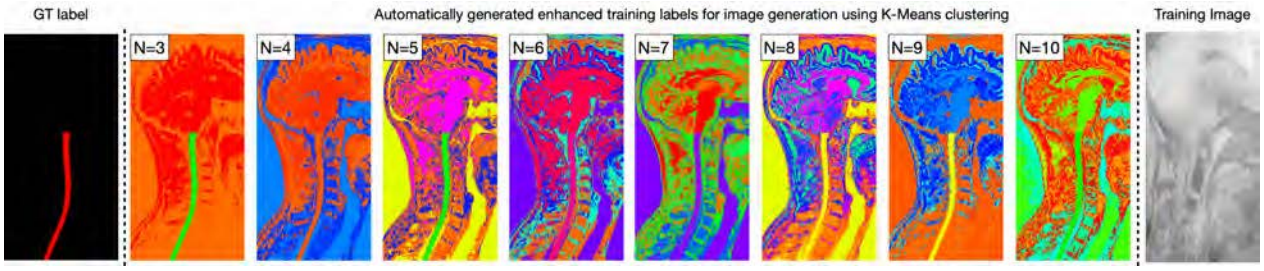

Figure B.9 Intensity-based K-Means clustering for automatic generation of labels outside the spinal cord (GT label). In all the enhanced labels, the spinal cord label value is fixed to 1 and the rest of the image is clustered between 3-10 clusters. One of these labels is randomly picked for image generation resulting in the training image.

For a fair comparison with our proposed model `soft_all`, we updated the spatial deformation parameters, namely, flipping, shearing and bias field to lie close to the range with which our model was trained. The default activation function was also changed from `elu` to `relu`. The model was trained for 50 epochs with 2500 steps per epoch with a batch size of 2.

Figure B.10 shows SynthSeg predictions on T1w, T2w, and T2star scans from the same

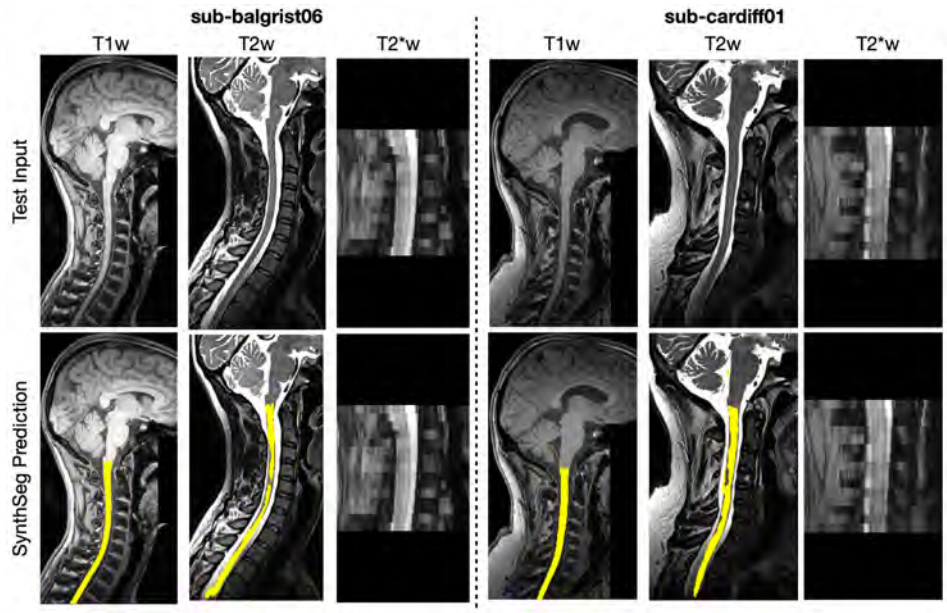

Figure B.10 SynthSeg predictions on T1w, T2w, and T2star contrasts for a given healthy subject. While the prediction on T1w scan is excellent, SynthSeg failed to properly segment the spinal cord on T2w (clear under-segmentation) and T2*w contrasts (no output segmentation).

subject. SynthSeg produced a complete prediction on the T1w scan, partial segmentation on the T2w scan and no segmentation for the T2star scan. It is worth noting that the original SynthSeg was evaluated only on T1w, T2w, FLAIR and proton density scans, however, contrasts such as T2star are common in spinal cord imaging and are used for segmenting the gray matter. Hence, the evaluating SynthSeg on contrasts other than the ones shown in the original paper is crucial. Incomplete/no predictions suggest that uniformity in the field-of-view (FoV) that the training labels cover might be playing an important role. In other words, all brain scans may contain all the standard labels however they are acquired, but, spinal cord scans such as T2star and DWI do not cover all the vertebral levels. Further, the range of Gaussian Mixture Model parameters used for generating synthetic scans might not be robust enough to model the contrasts seen in spinal cord imaging, thus leading to poor performance outside of T1w/T2w contrasts. More details on the re-training procedure along with a few synthetic scans and output segmentations on a few more contrasts can be found here.

APPENDIX C

Workshop paper: A comparative study of ConvNets and Vision transformers for contrast-agnostic spinal cord segmentation

Introduction

The cross-sectional area (CSA) of the spinal cord (SC) is an important biomarker for assessing cord compression and atrophy in neurological diseases such as multiple sclerosis (MS). However, the existing methods for SC segmentation have a *key* limitation: the predicted segmentation depends on the type of input MRI contrast and its acquisition parameters, resulting in different SC CSA for different MRI contrasts [17]. Furthermore, such methods are dominated by CNN-based approaches [17, 18], suggesting a gap in the literature to evaluate other deep learning (DL) architectures, namely, ConvNeXt and ViT-based approaches for SC segmentation.

There exist several studies comparing the performance of vision transformers (ViTs) and CNNs primarily focusing on classification tasks [276–278]. However, the conclusions from these studies are not generalizable as they are heavily dependent on the: (i) type of task (classification/segmentation/detection), (ii) input modality (i.e. digital pathology/natural/medical images), (iii) dataset sizes, and (iv) initialization strategy (i.e. from scratch/pretrained). Therefore, in this work, given a small dataset of spinal cord MRIs with multiple contrasts, we compared the performance of the modern DL architectures (namely, CNNs, ConvNeXT, and ViTs) for automatic spinal cord segmentation and evaluated their ability to achieve contrast-agnostic cord segmentation.

Materials and Methods

Dataset We used the open-access Spine Generic Public Database[1] consisting of healthy participants scanned on 3T MRI scanners across 42 sites. It consists of 6 MRI contrasts (T1w, T2w, T2*w, MT-on, GRE-T1w, and DWI) with both isotropic ($\{0.8, 1\}$ mm^3) and anisotropic ($\{0.5, 0.9\} \times \{0.5, 0.9\} \times \{3, 5\}$ mm^3) resolutions for each participant. This dataset presents a diverse set of MRI contrasts per participant to evaluate the models' contrast-agnostic segmentation capabilities. The final dataset ($n = 243$) was split according to 60%/20%/20% train/val/test splits, resulting in 145/49/49 participants with 870/294/294 3D volumes.

[1] https://github.com/spine-generic/data-multi-subject/releases/tag/r20231212

Preprocessing To eliminate the differences in CSA within the GT masks across contrasts, our preprocessing strategy produced a unique, *soft* GT mask averaged across all MRI contrasts (please see Section 2.2 of [18] for details).

Training Protocol We followed the SoftSeg [15] training strategy, treating the segmentation task as a regression problem, where, (i) we *do not* binarize the inputs fed to the model after data-augmentation and, (ii) instead of DiceLoss [134], we use adaptive wing loss [279] which penalizes higher errors at the SC boundary. For a given subject, each contrast is treated as an independent input during training as opposed to concatenating all 6 contrasts as channels in a multi-modal input.

Experiments and Results

Models and Training Details We compared 7 models spanning 3 different classes of DL architectures. For CNNs, we compared DeepSeg 2D [17], nnUNet [173], and the contrast-agnostic model [18]. The DeepSeg 2D and contrast-agnostic models are both accessible via the open-source Spinal Cord Toolbox [86] using the commands `sct_deepseg_sc` and `sct_deepseg -task seg_sc_contrast_agnostic` respectively. Then, we trained MedNeXt [280], a state-of-the-art ConvNeXt model designed for 3D medical images. Lastly, among ViTs, we compared UNETR [257], SwinUNETR [281], and an open-source, pretrained SwinUNETR[2]. Except for the pretrained model (which we fine-tuned on our dataset), all other models were trained from scratch for 200 epochs using Adam optimizer with a learning rate of 0.001 and a batch size of 2.

Evaluation For quantitative assessment of the variability of CSA across contrasts, we used the CSA averaged across individual slices from the C2-C3 vertebral levels as the primary metric. Specifically, we (i) obtained the model predictions for each contrast, (ii) computed the absolute error between the CSA of the model prediction and the GT for each contrast, and finally, (iii) averaged the CSA errors across 6 contrasts for each subject (shown as one scatter point in the violin plot). The lower the absolute CSA error (in mm^2), the better, as, in theory, the spinal cord CSA value should *not* vary substantially across MRI contrasts for a given participant.

Results Figure C.1 shows the absolute CSA error across contrasts for each test participants across all models. Among the CNN-based models, the `contrast-agnostic` model achieved

[2]https://github.com/Project-MONAI/tutorials/self_supervised_pretraining/swinunetr_pretrained

the lowest CSA error, with MedNeXt following closely. Interestingly, among the transformer models, UNETR showed the highest CSA error among all the models, while the SwinUNETR models showed similar performance as the CNNs.

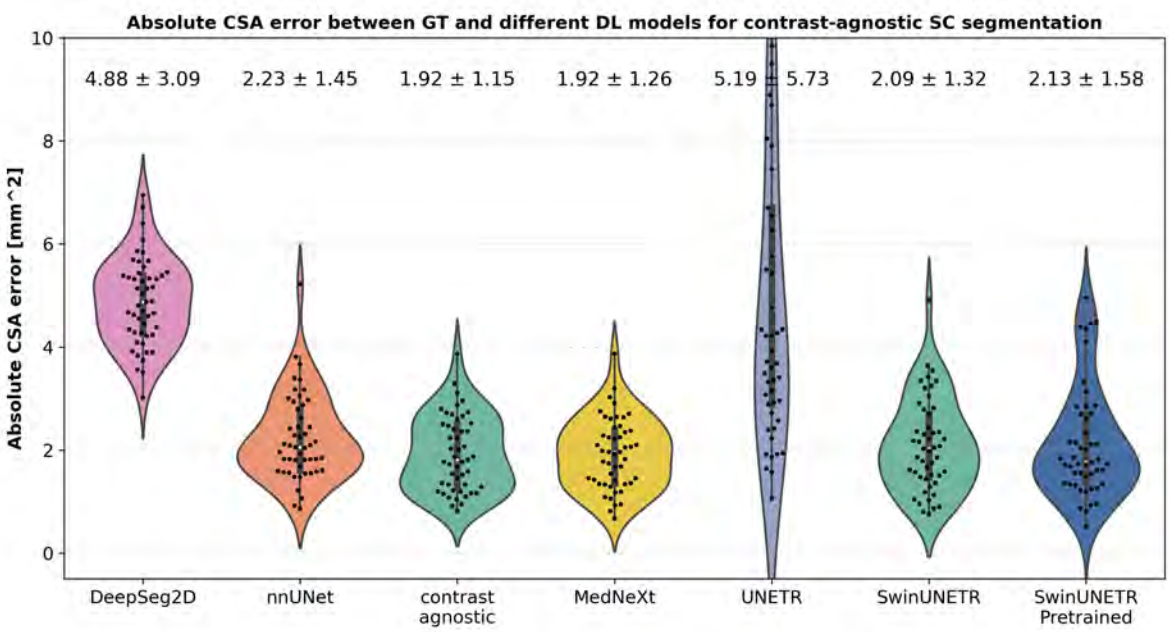

Figure C.1 Absolute CSA error between the GT and predictions averaged across all 6 MRI contrasts for each model. Scatter plots within each violin show the CSA error averaged across all contrasts for a given participant. White triangle marker shows the mean CSA error across test participants.

Discussion and Conclusion

In this study, we performed a preliminary analysis of the performance of different classes of DL architectures for the specific task of contrast-agnostic SC segmentation. Overall, given a fixed dataset size, the CNN-based methods produce more robust SC segmentations across MRI contrasts. UNETR, which processes fixed-resolution 3D patches of size $16 \times 16 \times 16$ as 1D sequence of tokens performs the worst, suggesting that weak inductive biases in pure transformer-based encoders can be a major limiting factor for segmentation tasks. Hierarchical ViTs such as SwinUNETR that bring back CNN-based priors (e.g. learning hierarchical representations via pooling and window-based local receptive fields, etc.) while using transformer blocks perform similar to CNNs. Future work aims at increasing the dataset size to include more contrasts and pathological images (such as MS) and comparing the performance of CNNs and SwinUNETR models at scale.

APPENDIX D

Workshop paper: Segmentation of multiple sclerosis lesions across hospitals: Learn continually or train from scratch?

Abstract

Segmentation of Multiple Sclerosis (MS) lesions is a challenging problem. Several deep-learning-based methods have been proposed in recent years. However, most methods tend to be *static*, that is, a single model trained on a large, specialized dataset, which does not generalize well. Instead, the model should learn across datasets arriving sequentially from different hospitals by building upon the characteristics of lesions in a *continual* manner. In this regard, we explore experience replay, a well-known continual learning method, in the context of MS lesion segmentation across multi-contrast data from 8 different hospitals. Our experiments show that replay is able to achieve *positive backward transfer* and reduce catastrophic forgetting compared to sequential fine-tuning. Furthermore, replay outperforms multi-domain training, thereby emerging as a promising solution for the segmentation of MS lesions. The code is open-source and available at this link.

Introduction

Multiple Sclerosis (MS) is a chronic, neurodegenerative disease of the central nervous system. Lesion segmentation from magnetic resonance images (MRI) serves as an important biomarker in measuring disease activity in MS patients. However, manual segmentation of MS lesions is a tedious process, hence motivating the need for automated tools for segmentation. Several deep-learning (DL) based methods have been proposed in the past few years [282, 283]. They tend to be trained in a *static* manner - all the datasets are pooled, jointly preprocessed, shuffled (to ensure they are independent and identically distributed, IID) and then fed to the DL models. While this has its benefits, it does not represent a realistic scenario. First, it is difficult to pool datasets from multiple hospitals with increasing privacy concerns. Second, since MS is a chronic disease, one would imagine a scenario where a DL model, like humans, engages in continual learning (CL) [87] and builds upon the lesion characteristics from different centers when presented sequentially. However, this se-

quential knowledge acquisition presents a major problem in DL models known as *catastrophic forgetting* [89].

Previous works in CL for medical imaging have used regularization-based [284, 285] and memory-based methods [286] for tackling catastrophic forgetting. In this work, we formalize the MS lesion segmentation across multiple hospitals as a domain-incremental learning problem [287], where the task remains unique (*i.e.* segmentation of lesions) but the model is sequentially presented with new domains (*i.e.*, data from different hospitals). Four types of experiments are performed as shown in Figure D.1 - *single-domain*, *multi-domain*, *sequential fine-tuning*, and *experience replay*. Our results show that replay helps in reducing catastrophic forgetting and achieves *positive backward transfer*, that is, the segmentation performance on data seen earlier improves as the model continues to learn sequentially. Furthermore, we show that replay outperforms multi-domain training as more data arrive sequentially, thereby suggesting that the CL is a better long-term solution than re-training the model from scratch on a large, curated dataset.

Experience replay for brain MS lesion segmentation

Replay (or, rehearsal) presents a straightforward way to prevent catastrophic forgetting and improve the performance on new domains. Let x denote the patches of the 3D volumes, y be the corresponding labels, f_θ denote the neural network with parameters θ, and $\mathcal{D}$ denote the joint dataset. In a standard IID training regime, the loss $\mathcal{L}$ is given by Equation D.1.

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim D}\left[\ell(f_\theta(x), y)\right] \tag{D.1}$$

$$\mathcal{L}' = \mathbb{E}_{(x,y)\sim D_k}\left[\ell(f_\theta(x), y)\right] + \mathbb{E}_{(x,y)\sim \mathcal{M}}\left[\ell(f_\theta(x), y)\right] \tag{D.2}$$

In this work, we use the simplest form of experience replay wherein training data from previously encountered domains[1] are stored in a memory buffer and interleaved with the current domain's training data. Particularly, the dataset $\mathcal{D}$ is divided into 8 different domains $(\mathcal{D}_1, \mathcal{D}_2, \ldots \mathcal{D}_8)$. The model is trained sequentially on one dataset $\mathcal{D}_k$ at a time. For each dataset $\mathcal{D}_k$ $(1 \leq k < 8)$, we store upto 20 image-label pairs (depending on the dataset size) in a memory buffer $\mathcal{M}$ and merge them with the training data of the current domain. The updated loss term $\mathcal{L}'$ is given by Equation D.2. Due to unconstrained access to the multi-center data, the model is tested on all the remaining centers once it has been trained on one center. We use the Dice Loss [134] as the loss function ℓ.

[1]We use *domains* and *centers* interchangeably. A domain is essentially a center (i.e. a hospital) that holds/provides the data. Hence, data from each new center is treated as a different domain.

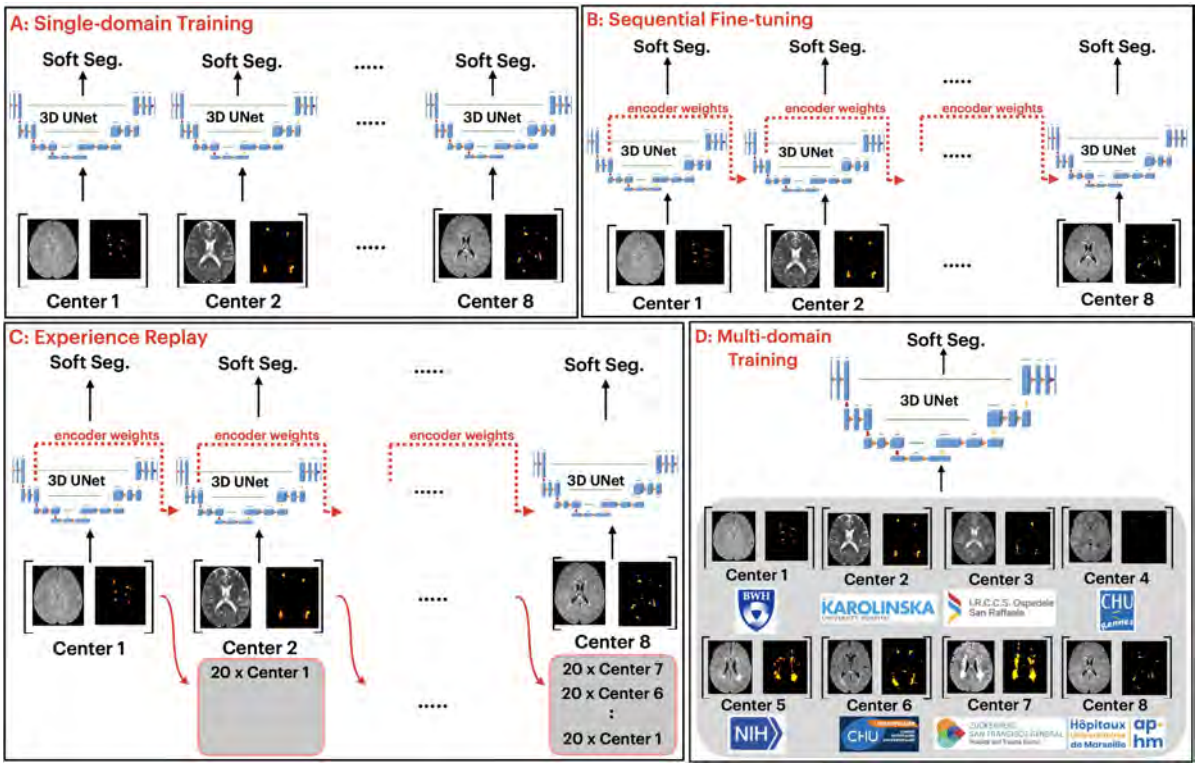

Figure D.1 Overview of our methods. Four experiments were performed - A: *Single-domain training*: a model is trained individually on each center. B: *Sequential fine-tuning*: after training the model on center n, the pre-trained encoder weights are loaded for center $n+1$ (red dashed arrows). C: *Experience replay*: in addition to fine-tuning (as in B) upto 20 samples per each center are stored in the memory buffer (in gray). D: *Multi-domain training*: data from all centers are pooled and a joint model is trained.

SoftSeg To account for the partial volume effects at the lesion boundaries, we use soft ground truth labels in our training procedure. In addition to mitigating the partial volume effects, soft segmentations [15] were shown to generalize better and to reduce model-overconfidence. In this regard, the notable changes include: (i) bypassing the binarization step after data preprocessing and augmentation, hence keeping the labels between $[0, 1]$, and (ii) using *normalized ReLU* as the final activation function.

Experiments and Results

Data We used the brain MRI datasets described in Kerbrat et al. [288] containing 290 subjects from 8 different centers. We denote the centers with the following abbreviations along with their number of subjects: BWH: $n = 80$, Karo: $n = 51$, Milan: $n = 47$, Rennes: $n = 51$, NIH: $n = 28$, Montp: $n = 13$, UCSF: $n = 12$, and AMU: $n = 8$. Six out of

eight centers used 3D FLAIR scans, center *Karo* used both 3D FLAIR and T2-weighted (T2w) scans, and center *Milan* used only T2w scans. The data were pre-processed using the publicly-available Spinal Cord Toolbox [?] and Anima Toolbox [176]. The intra-subject registration between the T_2 and the FLAIR images was achieved using rigid transformations and subsequently registered to the ICBM template space. All the 3D MRI images were resampled to an isotropic 1mm resolution. We refer the reader to [288] for more details.

Experiments A three-layer 3D UNet [128] with residual connections was used. The data were split according to the 80/20 train/test ratio. For fine-tuning and replay experiments, the model observed each domain in sequence. The ordering of domains was randomly shuffled with 9 different seeds.

Evaluation Metrics On a held-out test set, we used the Dice coefficient to evaluate the quality of the lesion segmentations and computed the backward transfer (BWT) metric [289] to evaluate the CL capabilities of our model. Concretely, BWT quantifies the influence that training on center n has on the performance on a previous center $k < n$. Hence, a positive BWT occurs when the Dice score on a center $k < n$ *increases* after training on center n and vice-versa for a negative BWT.

Results Figure D.2A shows the mean zero-shot test performance over 2 random sequences. When learning to transfer knowledge across FLAIR $\rightarrow$ T2 contrasts, we observed large drops in Dice scores on center *milan* as a consequence of catastrophic forgetting. On the other hand, replay performs better in this case and also exceeds the performance on multi-domain training. More importantly, Figure D.2B shows that over 9 random sequences of the domains, replay improves the segmentation performance over fine-tuning **and** multi-domain training as more data arrive. Thus, the proposed CL approach is capable of surpassing the implicit upper-bound defined by multi-domain (IID) training. In Figure D.3, we show the soft segmentations (ranging from $[0, 1]$) obtained from fine-tuning and experience replay on a test sample from the *milan* center. Fine-tuning incorrectly segments the periventricular region of the brain as lesions, whereas replay results in a better segmentation, while also providing a measure of uncertainty at the boundaries.

In Table D.1, we report the BWT in terms of the test Dice scores on a fixed, descending order of the domains (defined as per the number of subjects). Large negative BWT was observed especially with centers *karo* and *milan*, implying catastrophic forgetting. On the other hand, not only does replay improve performance on these centers, it also achieves a *positive BWT*, implying that training on the rest of the domains is indeed beneficial for these 2 domains.

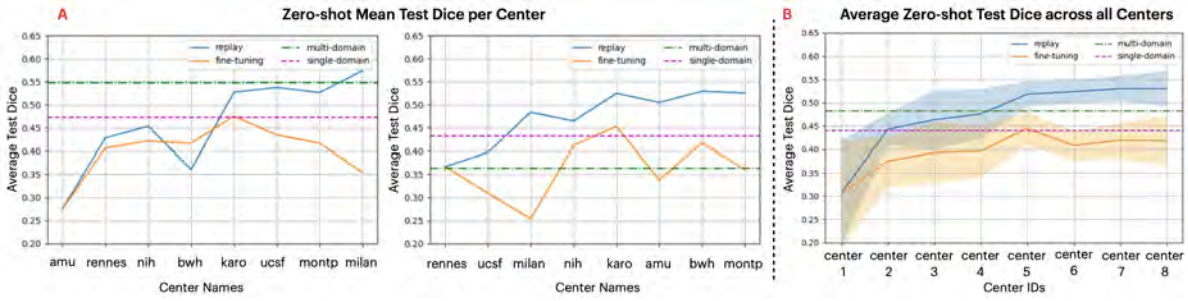

Figure D.2 Zero-shot (ZS) Test Dice scores with different random sequences of domains. A: ZS Test Dice scores with 2 random domain sequences. B: ZS Test Dice scores averaged across 9 randomly shuffled domain sequences.

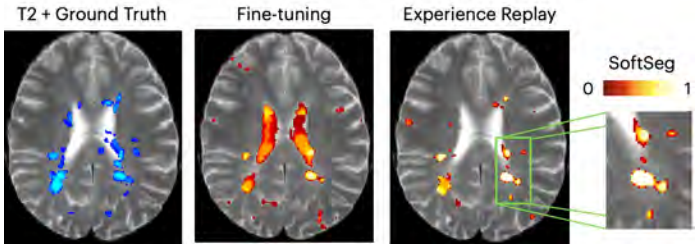

Figure D.3 Qualitative results on a test sample from *milan* center. Replay obtains better soft segmentations compared to fine-tuning.

Table D.1 BWT over descending order of domains (averaged across 9 seeds)

Center ↓	**Backward Transfer (BWT)**	
Method →	**Fine-tuning**	**Replay**
BWH	-0.07	**0.005**
Karo	-0.171	**0.037**
Milan	-0.284	**0.001**
Rennes	-0.083	-0.012
NIH	-0.119	-0.002
Montp	-0.061	**0.032**
UCSF	-0.061	-0.008
AMU	0.0	0.0

Conclusion

This work presented a case for continual learning using experience replay for the segmentation of MS lesions. Qualitative and quantitative results show that replay performs better than sequential fine-tuning in general, and especially when learning across contrasts FLAIR ↔ T2. More importantly, it also outperforms multi-domain (IID) training as the data continue to arrive. Thus, storing a few samples per domain and rehearsing them regularly can improve performance over the long-term, instead of re-training the model from scratch with each additional domain, which can be impractical.

APPENDIX E

Supplementary material: Monitoring morphometric drift in lifelong learning segmentation of the spinal cord

Full list of authors' affiliations

Authors Enamundram Naga Karthik[1,2], Sandrine Bédard[1], Jan Valošek[1,2,3,4], Christoph S. Aigner[5,6], Elise Bannier[7], Josef Bednařík[8,9], Virginie Callot[10,11], Anna Combes[12,13], Armin Curt[14], Gergely David[14,15], Falk Eippert[16], Lynn Farner[14], Michael G Fehlings[17,18], Patrick Freund[14,19,20], Tobias Granberg[21,22], Cristina Granziera[23], RHSCIR Network Imaging Group[24], Ulrike Horn[16], Tomáš Horák[8,9], Suzanne Humphreys[24], Markus Hupp[14], Anne Kerbrat[25,26], Nawal Kinany[27,28], Shannon Kolind[29], Petr Kudlička[9,30], Anna Lebret[14], Lisa Eunyoung Lee[31], Caterina Mainero[32], Allan R. Martin[33], Megan McGrath[13], Govind Nair[34], Kristin P. O'Grady[13], Jiwon Oh[35], Russell Ouellette[21,22], Nikolai Pfender[14], Dario Pfyffer[36,14], Pierre-François Pradat[37], Alexandre Prat[38], Emanuele Pravatà[39,40], Daniel S. Reich[41], Ilaria Ricchi[27,28], Naama Rotem-Kohavi[24], Simon Schading-Sassenhausen[14], Maryam Seif[14,19], Andrew Smith[42], Seth A Smith[13], Grace Sweeney[13], Roger Tam[43], Anthony Traboulsee[44], Constantina Andrada Treaba[32], Charidimos Tsagkas[41,23], Zachary Vavasour[43], Dimitri Van De Ville[27,28], Kenneth Arnold Weber II[45], Sarath Chandar[2,46], Julien Cohen-Adad[1,2,47,48]

Affiliations

1. NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montréal, Montréal, Québec, Canada
2. Mila - Québec Artificial Intelligence Institute, Montréal, Québec, Canada
3. Department of Neurosurgery, Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc, Czechia
4. Department of Neurology, Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc, Czechia
5. Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany
6. Max Planck Research Group MR Physics, Max Planck Institute for Human Development, Berlin, Germany
7. Department of Neuroradiology, Rennes University Hospital, Rennes, France
8. Department of Neurology, University Hospital Brno, Brno, Czechia
9. Faculty of Medicine, Masaryk University, Brno, Czechia
10. Aix-Marseille Univ, CNRS, CRMBM, Marseille, France
11. APHM, CHU Timone, CEMEREM, Marseille, France

12. NMR Research Unit, Queen Square Multiple Sclerosis Centre, UCL Queen Square Institute of Neurology, University College London, London, UK
13. Vanderbilt University Institute of Imaging Science, Vanderbilt University Medical Center, USA
14. Spinal Cord Injury Center, Balgrist University Hospital, University of Zurich, Zurich, Switzerland
15. Department of Neuro-Urology, Balgrist University Hospital, University of Zurich, Zurich, Switzerland
16. Max Planck Research Group Pain Perception, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
17. Division of Neurosurgery and Spine Program, Department of Surgery, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada
18. Division of Neurosurgery, Krembil Neuroscience Centre, UHN, Toronto, ON, Canada
19. Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
20. Wellcome Trust Centre for Neuroimaging, Queen Square Institute of Neurology, University College London, London, United Kingdom
21. Department of Neuroradiology, Karolinska University Hospital, Stockholm, Sweden
22. Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
23. Translational Imaging in Neurology (ThINk), Department of Biomedical Engineering, Faculty of Medicine, Basel, Switzerland
24. Praxis Spinal Cord Institute, Vancouver, BC, Canada
25. EMPENN Research Team, IRISA, CNRS-INSERM-INRIA, Rennes Université, Rennes, France
26. Neurology Department, Rennes University Hospital, Rennes, France
27. Neuro-X Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Geneva, Switzerland
28. Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland
29. Division of Neurology, Department of Medicine and the Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, BC, Canada
30. Multimodal and Functional Imaging Laboratory, Central European Institute of Technology, Czechia
31. Institute of Medical Science, University of Toronto, Toronto, ON, Canada
32. Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, USA; Harvard Medical School, Boston, MA, USA
33. Department of Neurosurgery, University of California Davis, Davis, CA, USA
34. qMRI Core Facility, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA
35. Barlo MS Centre, Division of Neurology, Department of Medicine, St. Michael's Hospital, Canada
36. Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Palo Alto, CA, USA
37. Department of Neurology, Pitie-Salpetriere Hospital, Paris, France
38. Department of Neuroscience, Université de Montréal, Montréal, QC, Canada
39. Department of Neuroradiology, Neurocenter of Southern Switzerland, Lugano, Switzerland
40. Department of Neuroscience, Imaging and Clinical Sciences, Università G. d'Annunzio, Chieti, Italy
41. Translational Neuroradiology Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA
42. Department of Physical Medicine and Rehabilitation, University of Colorado School of Medicine, Aurora, CO, USA

43. School of Biomedical Engineering, Department of Radiology, The University of British Columbia, Vancouver, BC, Canada
44. Department of Medicine, Division of Neurology, University of British Columbia, BC, Canada
45. Division of Pain Medicine, Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Palo Alto, CA, USA
46. Canada CIFAR AI Chair
47. Functional Neuroimaging Unit, CRIUGM, Université de Montréal, Montreal, QC, Canada
48. Centre de Recherche du CHU Sainte-Justine, Université de Montréal, Montréal, QC, Canada

Dataset characteristics

Table E.1 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 × 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]	–	–	–	–

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 [16], 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` [17], `contrast_agnostic_v2.0` [18], and the proposed `contrast_agnostic_v3.0`, for contrasts and sites. In all comparisons, the spinal

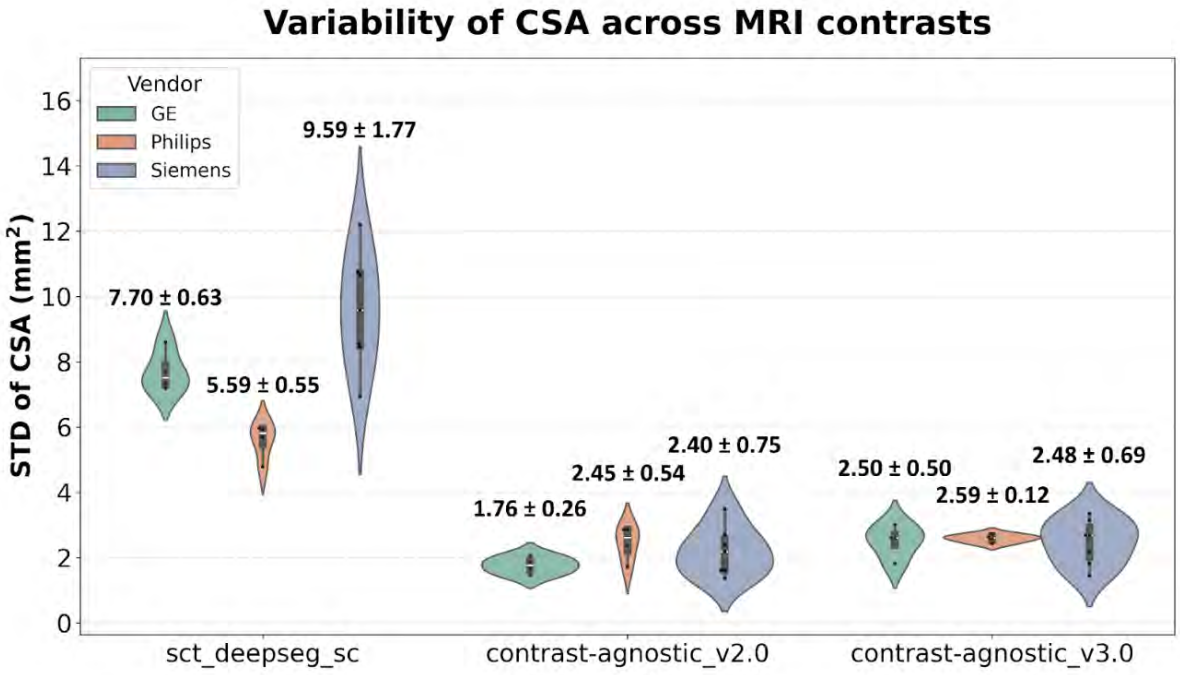

Figure E.1 Variability of spinal cord CSA across contrasts separated per vendor for segmentations generated with `sct_deepseg_sc` [17], `contrast_agnostic_v2.0` [18] and `contrast_agnostic_v3.0` (proposed) segmentation and contrast-agnostic of the same participant scanned across 15 different MRI sites. Each dot represents one site; mean and standard deviation are presented above.

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 Figure E.1 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`.

CSA variability with recursively generated labels

Figure E.2 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.

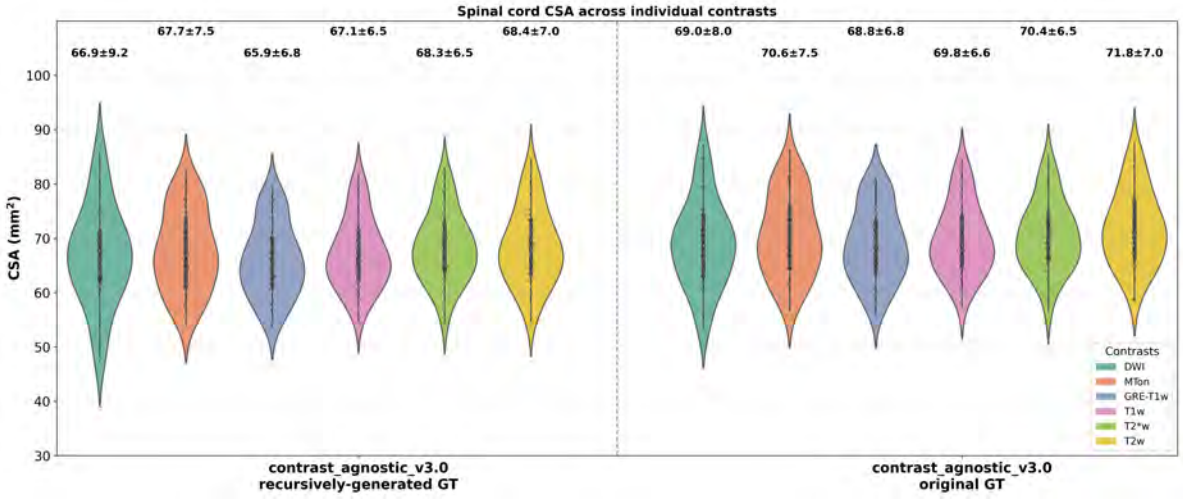

Figure E.2 Variability in spinal cord CSA across 6 contrasts on a test set of healthy participants ($n = 49$) compared between the models trained with the: (i) original distribution of GT masks created from a mix of manual annotations and automatic segmentation methods, and (ii) GT masks regenerated 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.