npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

Article in Press

https://doi.org/10.1038/s41746-026-03403-w

A deep learning framework for automated cervical spinal cord vertebral labeling and morphometric extraction from brain MRI with large-scale multicenter validation

Check for updates

A list of authors and their affiliations appears at the end of the paper

This study aimed to develop and validate a fully automated pipeline for vertebral level identification (C1–C3) of the upper cervical spinal cord (SC) in T1-weighted brain MRI with a quality control module to assess the generated masks. MRI data from 998 controls and 879 patients with hereditary ataxias (32 sites; age 5–85 years; 46% male) were used to train deep learning models (nnU-Net, U-Net with deep supervision, SwinUNETR) for automatic vertebral labeling of C1–C3 levels. A secondary model based on VGG, ResNet, and EfficientNet architectures was trained to identify vertebral labeling errors. The pipeline also generates image quality metrics including signal-to-noise ratio, contrast-to-noise ratio, coefficient of joint variation, and entropy focus criterion. External validation was performed using 2,737 MRIs from four public datasets, and clinical utility was demonstrated in a longitudinal cohort of 51 patients with multiple sclerosis. The nnU-Net achieved the highest performance (Dice: C1 = 0.85, C2 = 0.89, C3 = 0.89), labeling 95.8% of mislabeled cases from the Spinal Cord Toolbox. The VGG-based classifier achieved 97.9% accuracy and 98.7% sensitivity. External validation showed labeling accuracies exceeding 92% in four of five datasets and sensitivity to longitudinal SC atrophy in multiple sclerosis. This pipeline enables scalable, reliable vertebral labeling and morphometric extraction of the upper cervical SC for multicenter studies and supports biomarker development in neurodegenerative diseases.

The spinal cord (SC) is a crucial structure of the central nervous system, responsible for conducting signals between the brain and the body, playing a vital role in motor and sensory functions[1]. Structural changes in the SC, such as atrophy or deformities, are inherent in many neurological disorders[1]. Technical advances have enabled increasingly detailed acquisition and quantitative processing of SC structure using MRI[2]. This includes the ability to assess SC structure using standard brain 3D T1-weighted (T1w) MRI with coverage of the upper cervical SC. This approach has generated promising biomarkers in various neurodegenerative diseases, including amyotrophic lateral sclerosis (ALS), multiple sclerosis (MS), and degenerative ataxias[3–6]. The use of a single MRI to assess both brain and upper SC structure provides efficient translational potential for neurological and radiological outpatient practice and research applications[7]. It also enables the evaluation of upper SC in existing large public datasets (e.g., UK Biobank)[8] and research consortia (e.g., ENIGMA)[3,4] where brain MRI is ubiquitous, but SC-specific acquisitions are not available.

Accurate, reproducible, and generalizable automated quantification of SC structure is challenging due to the complexity and variability of SC structure and surrounding vertebrae and soft tissues, in-scanner positioning (i.e., neck curvature), and MR signal characteristics. These challenges are magnified in datasets that include SC damage and/or pediatric subjects[9,10]. As a consequence, high rates of segmentation or quantification failures occur in clinical samples when using currently available automated algorithms, leading to substantial data loss and/or labor-intensive procedures to manually correct errors[3,4,9].

Recent advancements in deep learning offer promising solutions to address data generalizability and accuracy. Deep learning algorithms, particularly convolutional neural networks (CNNs), can learn complex patterns from large datasets, making them well-suited for the task of automated upper SC vertebral labeling[11]. By leveraging extensive training datasets, these models can account for image and anatomical variability, effectively adapting to the diverse morphologies and imaging conditions encountered

e-mail: thiago.jrezende@gmail.com

in clinical practice[12]. Moreover, the integration of deep learning with quality control (QC) processes can facilitate the identification of labeling errors and segmentation inaccuracies from upstream processing, leading to more reliable analyses and enhanced data integrity, especially in larger cohorts[13]. Indeed, regulatory agencies explicitly emphasize that QC is a necessary methodological safeguard to ensure robustness, scalability, and clinical applicability of imaging pipelines[14].

In this work, we generate an optimized deep-learning-based analytical tool to accurately label the vertebral levels of the upper SC in human brain 3D T1w MRI. The algorithm is trained using data from healthy individuals across the lifespan and individuals with qualitatively and quantitatively variable SC degeneration from the TRACK-FA[9], ENIGMA-Ataxia[3,4], and READISCA[15–17] research consortia. To achieve a fully automated and standardized tool that can be readily deployed for use in observational, natural history, and treatment trials, a QC module was also developed to provide quantitative metrics and automated visual inspection. This tool is publicly available for use in observational and clinical trials (https://github.com/art2mri/Enigma-SC).

Results

For this study, we enrolled a total of 4665 brain 3D T1w MRI, 924 in vertebral labeling training, 1396 specifically for training the automated visual QC (Table 1), and 2788 MRI volumes for external testing (Table S1). The age ranges from 5 to 85 years, and 2112 (45.3%) males, 2553 (54.7%) females, 1455 (31.2%) pediatric individuals, and 3210 (68.8%) adults. Healthy individuals account for 3,735 (80.0%) of the MRI volumes followed by 354 (7.6%) Friedreich's Ataxia (FRDA), 119 (2.6%) spinocerebellar ataxia (SCA) type 1 (SCA1), 87 (1.9%) SCA2, 246 (5.3%) SCA3, 25 (0.5%) SCA6, 48 (1.0%) SCA7 and 51 (1.1%) MS.

Labeling task

For the vertebral labeling task, the nnU-Net model exhibited the highest performance across all evaluated metrics (Tables 2 and S2). Although no statistically significant differences between the evaluated architectures were observed (Table S2). The nnU-Net framework was therefore selected based on its recognized robustness, consistent performance, stable convergence behavior, and widespread adoption for segmentation tasks in medical imaging. The pipeline ran in approximately 32 s/image on an Nvidia RTX A6000 48 GB. Using CPU-only hardware, the estimated processing time was 2.3 h/image. The final trained model (nnU-Net) provided C2–C3 spinal cord metrics that were very highly correlated with outcomes generated by the reference toolbox (Spinal Cord Toolbox, SCT) for cross-sectional area (CSA) ($R = 0.84–0.95$) and more moderately for eccentricity (ECC) ($R = 0.58–0.68$) in both adults and children (age <18 years old) (Figs. 1 and 2). Poorer correspondence was observed at the C1 level in both cohorts (Figs. 1 and 2), especially in children, potentially due to substantial variability in the spatial localization of the SCT C1 vertebral level. This pattern is replicated when the data are divided into ataxic and healthy control cohorts (Figs. S1 and S2). This variability in SCT-derived metrics is markedly reduced when employing the nnU-Net model, indicating improved consistency and anatomical alignment across subjects. This is in line with the superior performance of nnU-Net model on the testing dataset (698 MRI volumes), which was able to correctly identify the vertebral levels in 95.8% of the previously failed cases (669 of 698 MRI volumes) (Table S3).

Automated quality control

The VGG-net architecture demonstrated the highest performance on the automated quality control task, achieving high accuracy, sensitivity, and specificity to detect both correct and uncorrected vertebral labeling with minimal false positive and false negative cases (Table 3). EfficientNet-B7 also showed strong performance (Table 3), while the ResNet model achieved slightly lower performance (Table 3). Despite the minimal variation in accuracy across models, a key difference was observed in convergence speed: ResNet and VGG-Net reached stable performance within approximately 7

Table 1 | Demographic data of the complete dataset of collected MRI volumes

	TRACK-FA		ENIGMA-Ataxia							READISCA		
	Control	FRDA	Control	FRDA	SCA1	SCA2	SCA3	SCA6	SCA7	Control	SCA1	SCA3
Subjects	54	134	930	220	96	87	183	25	48	14	23	63
Children (<18 years)	26	55	31	42	0	3	0	0	0	0	0	0
Age (mean ± SD)	20.1 ± 9.5	21.7 ± 9.8	36.8 ± 14.1	28.4 ± 12.8	44.4 ± 11.9	42.2 ± 13.9	46.7 ± 12.3	64.8 ± 10.0	47.9 ± 13.6	45.5 ± 7.2	41.0 ± 8.3	42.0 ± 8.2
Sex (M/F)	27/27	71/63	399/531	95/125	55/41	45/42	87/96	15/10	24/24	8/6	7/16	27/36

FRDA Friedreich's ataxia, *SCA* Spinocerebellar Ataxia type 1, 2, 3, 6, and 7, *SD* standard deviation, *M* males, *F* females.

Table 2 | Performance metrics of the trained deep learning models for vertebral labeling (C1–C3) in the internal validation dataset

Vertebral level	Model	Dice	IoU (Jaccard)	Precision	Recall	Hausdorff distance
C1	nnU-Net	**0.854 ± 0.029**	**0.749 ± 0.032**	**0.889 ± 0.033**	**0.829 ± 0.041**	**4.121 ± 0.659**
	U-Net DS	0.848 ± 0.037	0.736 ± 0.034	0.870 ± 0.029	0.812 ± 0.034	4.823 ± 0.431
	SwinUNETR	0.841 ± 0.031	0.726 ± 0.034	0.860 ± 0.027	0.823 ± 0.033	5.822 ± 0.729
C2	nnU-Net	**0.893 ± 0.021**	**0.807 ± 0.030**	**0.930 ± 0.029**	**0.875 ± 0.032**	**2.349 ± 0.414**
	U-Net DS	0.891 ± 0.034	0.803 ± 0.029	0.910 ± 0.028	0.864 ± 0.031	2.715 ± 0.498
	SwinUNETR	0.892 ± 0.029	0.805 ± 0.031	0.910 ± 0.028	0.873 ± 0.037	2.576 ± 0.663
C3	nnU-Net	**0.886 ± 0.024**	**0.796 ± 0.031**	**0.901 ± 0.027**	**0.890 ± 0.036**	**3.063 ± 0.532**
	U-Net DS	0.882 ± 0.031	0.778 ± 0.031	0.863 ± 0.028	0.873 ± 0.030	3.491 ± 0.557
	SwinUNETR	0.883 ± 0.025	0.779 ± 0.033	0.890 ± 0.023	0.886 ± 0.032	3.214 ± 0.590

U-Net DS U-Net with Deep Supervision, *IoU* Intersection over Union.

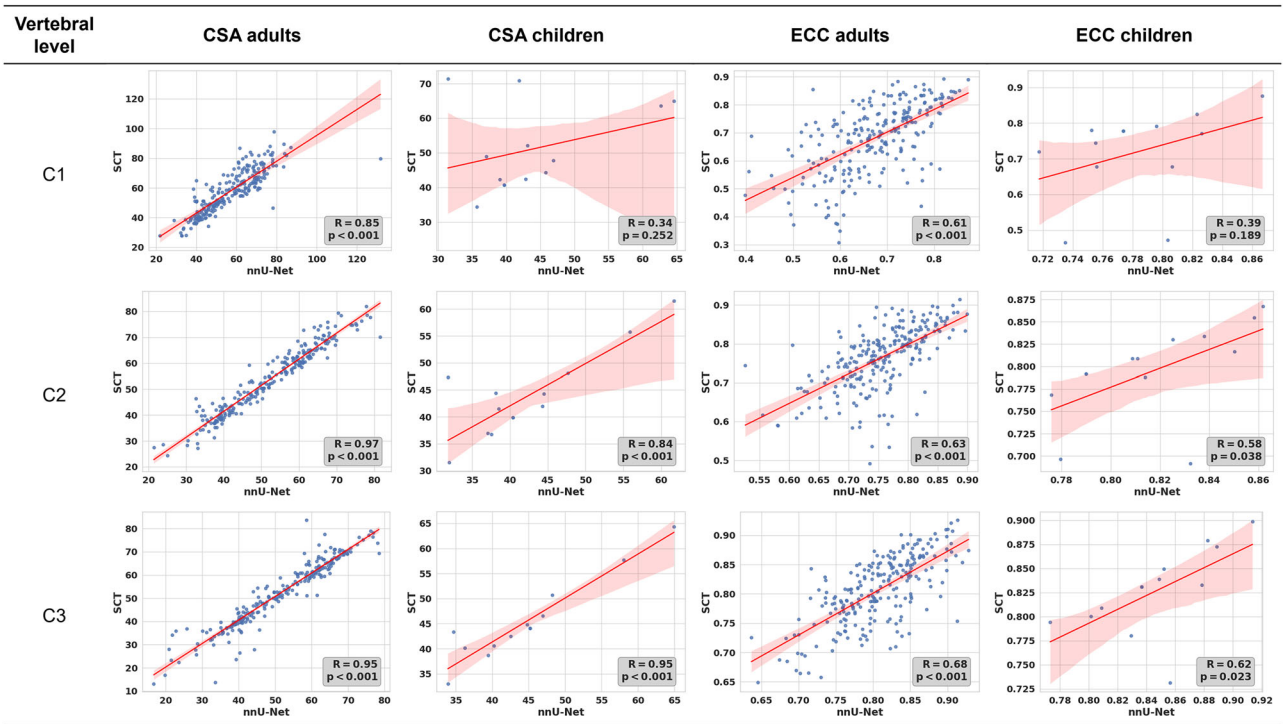


Fig. 1 | Correlation of spinal cord cross-sectional area and eccentricity from SCT and the proposed deep learning method. Scatterplots showing the linear relationship between spinal cord cross-sectional area (CSA in mm²) and eccentricity (ECC) measurements obtained using the Spinal Cord Toolbox (SCT) and those generated by the nnU-Net deep learning-based model in the test dataset ($N = 698$). Analyses were restricted to the subset of MRI volumes where both nnU-Net and SCT successfully generated valid masks ($N = 341$), excluding cases where labeling failed for either method ($N = 357$). Values represent the mean area across the first three cervical vertebral levels (C1–C3), stratified by age group: children (<18 years, $N = 12$) and adults (≥18 years, $N = 329$).

training epochs, whereas EfficientNet-B7 required around 20 epochs to converge.

External validation

The trained nnU-Net labeling model demonstrated robust performance when evaluated on an external dataset comprising 2788 MRI volumes from five independent public and private sources. In four out of the five datasets, the model achieved a vertebral labeling success rate of 92–99% (Table S4). The only exception was the C-MIND (Cincinnati MR Imaging of NeuroDevelopment)[18] dataset (78% accuracy), in which most MRI volumes included only a small portion of the SC, primarily restricted to the C1 region. This limited field of view (FoV) hindered the model's ability to reliably identify anatomical landmarks of the SC. In contrast, SCT failed on all C-MIND MRI volumes and had a labeling accuracy of 53–79% in the remaining four datasets (Table S4). Additionally, the nnU-Net model outputs correlated well with SCT measures, with the C1 vertebral level again showing the worst performance (Figs. S3 and S4). Outliers in the correlations were mostly ascribed to SCT misidentifications: in 49 out of 56 SCT or nnU-Net outliers, the nnU-Net generated correct vertebral level identifications; in 3 cases SCT provided correct results; and in 4 cases both methods failed (Fig. S5).

The VGG-based label error model had high accuracy (87–98%) in classifying correct and incorrect labeled images generated by the nnU-Net in four of the five external datasets (Table S5). The C-MIND (74% accuracy) and NIH (National Institutes of Health (NIH) MRI study of normal brain development)[19] (66% accuracy) datasets represent challenging MRI volumes for the model due to reduced FoV and image noise. Overall, these results demonstrate a tendency toward false negatives, which facilitates the

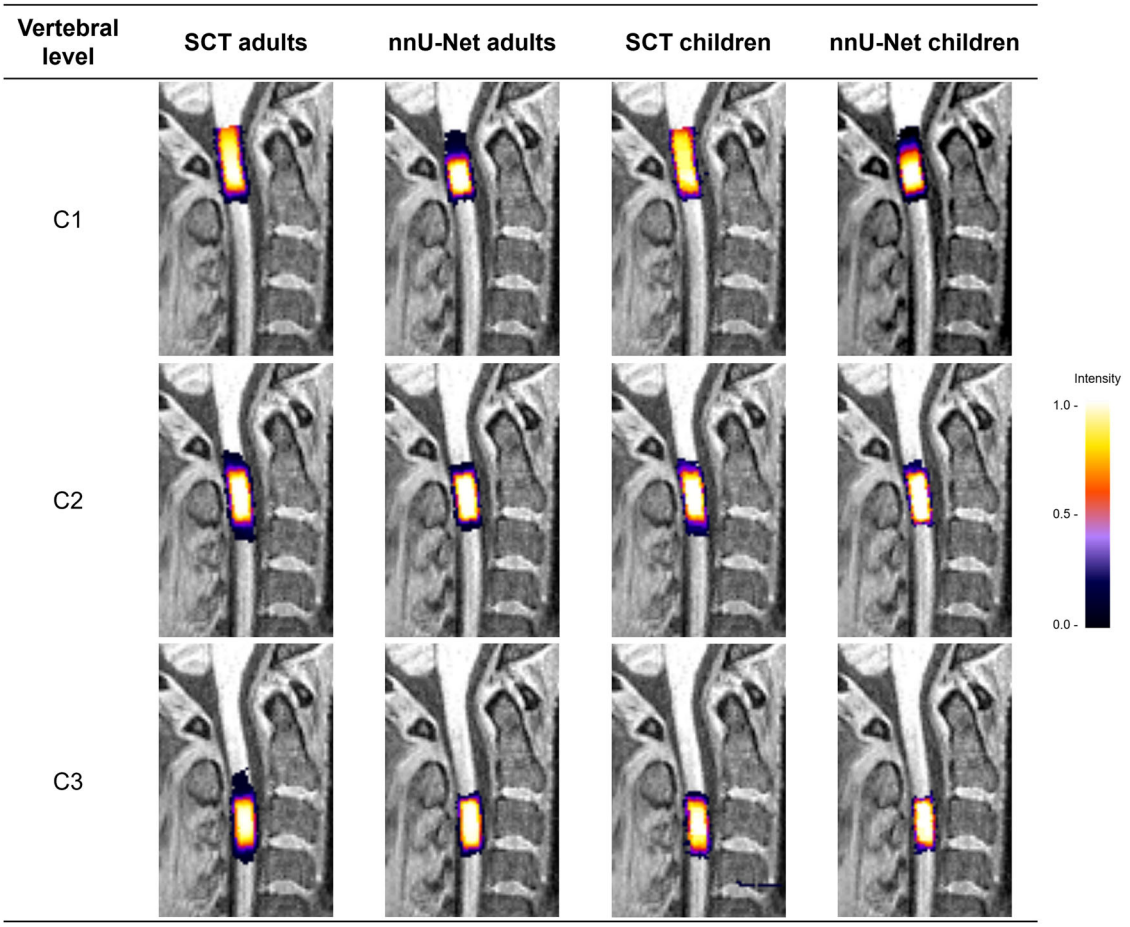


Fig. 2 | Spatial distribution of vertebral masks in a common template: SCT and our method results. Spatial overlap in the vertebral masks generated by the trained deep learning pipeline across all individuals in the test dataset ($N = 329$ adults, $N = 12$ children). Analyses were restricted to the subset of MRI volumes where both nnU-Net and SCT successfully generated valid masks ($N = 341$), excluding cases where labeling failed for either method ($N = 357$). Masks were generated at the individual level and spatially normalized to a common template. Signal intensity was normalized on a scale from 0 to 1 to represent the relative density of labeled regions. Areas with higher intensity indicate greater presence of the label in that location. C1, C2, C3 = first, second, and third cervical vertebral levels.

Table 3 | Performance metrics and training convergence epochs of the evaluated models on the automated quality control task (test dataset, *N* = 140)

Model	Accuracy (%)	Sensitivity	Specificity	TP	TN	FP	FN	Epochs to converge
VGG-net	97.9	**0.987**	**0.969**	75	62	2	1	~7
EfficientNet-B7	97.1	0.973	**0.969**	73	63	2	2	~20
ResNet	96.4	0.973	0.955	72	63	3	2	~7

TP true positive, *TN* true negative, *FP* false positive, *FN* false negative.

quality check step. Although not yet highly accurate, the system reduces the number of MRI volumes requiring visual inspection by accurately flagging correctly labeled images (Table S5). Only images flagged as mislabeled therefore require manual screening.

The quantitative image quality metrics were able to accurately differentiate different categories of SC imaging quality, except for the coefficient of joint variation (CJV), which did not show statistical significance between the groups ($p > 0.05$) (Fig. 3A, Table S6). The receiver operating characteristic (ROC) curve revealed signal-to-noise ratio (SNR) (area under the curve (AUC) = 0.998) and entropy focus criterion (EFC) (AUC = 0.924) as the best metrics to discriminate between acceptable (medium or high) versus low-quality MRI volumes (Fig. 3B, Table 4). Cut-off values for identifying low-quality data are SNR < 48 and EFC > 8.35×10^{-5} (Table 4). In the external testing datasets (excluding C-MIND), datasets with higher SNR and contrast-to-noise ratio (CNR) values and lower EFC values are more likely to be labeled as "correct" (Tables S5–S8). Notably, CJV was retained for completeness of evaluated image quality metrics but is not used for threshold-based exclusion decisions.

To further assess generalizability across sites and acquisition protocols, all image quality metrics were computed independently within each external dataset, again excluding C-MIND. The observed SNR, CNR, and EFC consistently reflected dataset-specific image characteristics, supporting their robustness across heterogeneous cohorts (Table S7). For example, NIH data acquired at 1.5 T and comprising predominantly pediatric subjects (Tables S1 and S9) exhibited lower SNR and CNR and higher EFC values (Table S7), consistent with lower field strength and increased motion-related artifacts. Similarly, MS datasets showed relatively preserved SNR and CNR but elevated EFC, consistent with increased motion and disease-related variability (Table S7). IXI (Information eXtraction from Images) (https://brain-development.org/ixi-dataset/) data, acquired predominantly

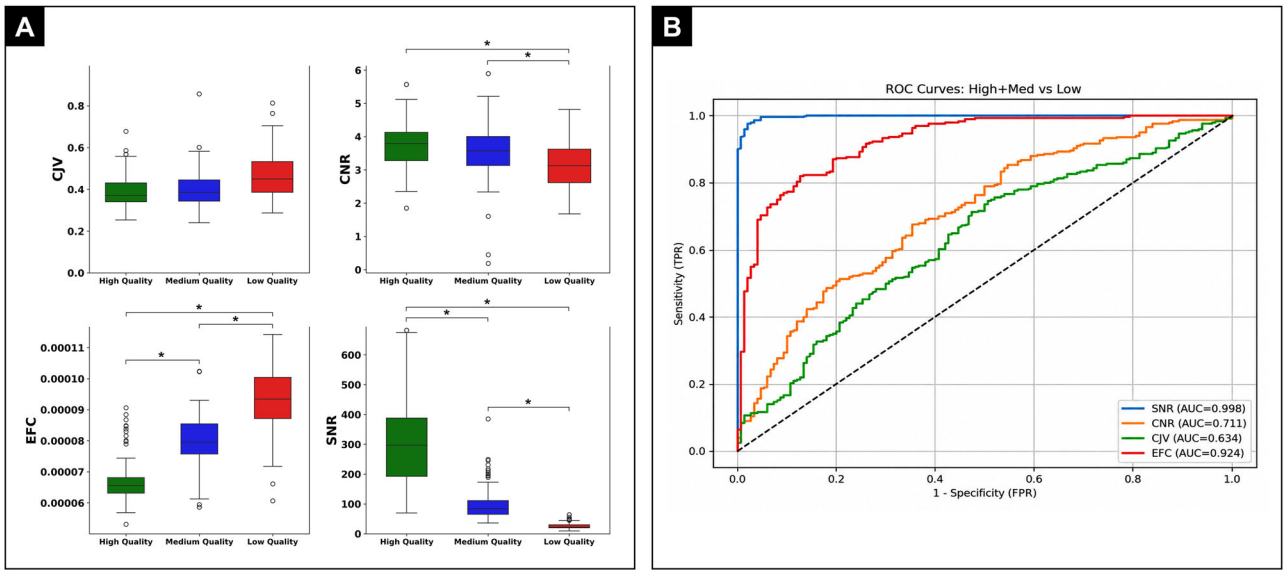


Fig. 3 | Quantitative MRI quality metrics relative to manual grading. Quantitative quality control metrics relative to manually graded MRI volumes. **A** Boxplots illustrate the pairwise-group comparison for each quantitative MRI quality metric. **B** Receiver Operating Characteristic (ROC) curves of classification performance in distinguishing good (medium- and high-quality MRI volumes) and low-quality MRI volumes. AUC area under the curve, $^{*}p < 0.05$. SNR signal-to-noise ratio, CNR contrast-to-noise ratio, CJV coefficient of joint variation, EFC entropy focus criterion.

Table 4 | Parameters obtained from receiver operating characteristic (ROC) curves to distinguish good (high- and medium-) quality images and low-quality images for each quantitative quality control metric, signal-to-noise ratio (SNR), contrast-to-noise ratio, coefficient of joint variation (CJV), and entropy focus criterion (EFC)

Metric	AUC	Cut-off	Sensitivity	Specificity
SNR	0.998	48.41	0.977	0.980
CNR	0.711	3.36	0.677	0.647
CJV	0.634	0.37	0.713	0.527
EFC ($\times 10^{-5}$)	0.924	8.35	0.820	0.873

The cut-off values represent the minimum value for a given parameter to be assigned as a good-quality image.
AUC area under the curve.

in adults across mixed 1.5 T and 3 T scanners (Tables S1 and S9), demonstrated generally favorable image quality metrics (Table S7), while Human Connectome Project Development (HCP-D) data showed reduced CNR and SNR in the spinal cord region, reflecting suboptimal coil coverage and background noise (Fig. S6 and Table S7).

Real-world application

To investigate SC-related MRI changes, we conducted an exploratory analysis of group-level trajectories in SC morphometric measures, CSA and eccentricity, of patients with MS and controls who were scanned longitudinally up to four times (mean follow-up 9.05 ± 4.67 years). Progressively decreasing CSA was evident at C1 and C2 vertebral levels in patients, while control participants remained relatively stable during follow-up (Fig. 4, Table S10). In contrast, eccentricity did not show significant change over time. The highest effect size was observed for C1 CSA, both within-group (standardized response means (SRM) = −0.73) and between-group (Cohen's *d* on slope = −0.70 for patients vs. controls).

Discussion

We developed a fully automated deep-learning-based toolbox to assess upper SC structure in T1w brain MRI. This represents a key clinical advantage, since brain MRI is more widely available than devoted spinal cord imaging and allows to gain information on the spine from brain scans. The toolbox provides accurate automated vertebral labeling in MRI volumes with upper spinal cord coverage including at least vertebral level C2, a QC report flagging likely low-quality MRI volumes or erroneous labels, and calculation of spinal cord cross-sectional area and eccentricity at each vertebral level. The pipeline exceeds the performance of a benchmark method and fully automates an existing semi-automated solution[4]. The algorithm was trained and tested using a wide variety of datasets and populations, including children, individuals with spinal cord degeneration, and diverse data augmentations. The final model showed robust performance during external validation and sensitivity to morphometric SC changes over time in a patient cohort. The pipeline is publicly available and readily scalable to large population datasets.

The optimized deep learning-based approach for vertebral level labeling addressed critical challenges that have affected the accuracy of existing methods[3–5,10]. In particular, the presence of anatomical alterations such as SC atrophy, scoliosis, and curvature often hinders the performance of spinal cord segmentation tools[3,4,9]. The nnU-Net approach employed here demonstrated enhanced robustness and adaptability across variable anatomy and image quality. Notably, the success of the proposed pipeline reflects the availability of a rich and diverse training dataset and a carefully defined labeling and quality control framework, rather than architectural novelty alone, as nnU-Net-based architectures are also employed within SCT. Current automated spinal cord processing pipelines show limited accuracy in vertebral level identification, with the most pronounced errors occurring at the C1 level due to its unique anatomy and lack of a well-defined intervertebral disc[10,20]. This often results in cranio–caudal mislocalization, leading to inadvertent inclusion of brainstem tissue or exclusion of spinal cord tissue, which artificially inflates or deflates CSA measurements[20]. Although template-based registration approaches reduce variability at the group level, vertebral correspondence at C1 remains particularly sensitive to inter-subject anatomical variability and pathological alterations[9,21]. By improving reliability in these challenging contexts, the model ensures greater consistency in SC analyses, which is important for studies aiming to detect subtle neurodegenerative changes or monitor disease progression[5,22].

A further strength of this pipeline is the incorporation of an automated QC system, which flags potential labeling errors and inconsistencies without requiring full visual verification of each case. Although the VGG-based label

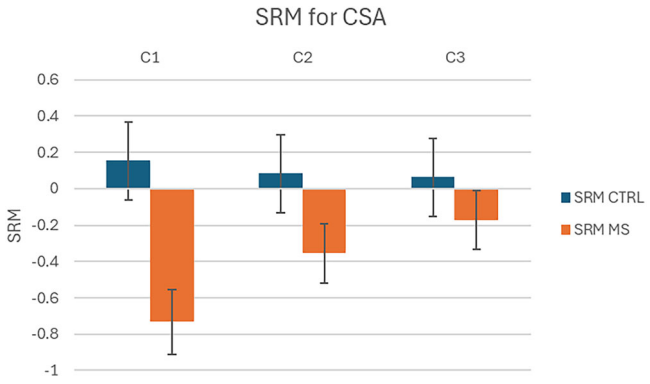


Fig. 4 | Standardized response means of cervical spinal cord cross-sectional area in multiple sclerosis and controls. Standardized response means (SRM) of cervical spinal cord cross-sectional area (CSA). Bar plots show the standardized response mean (SRM) for CSA at the C1, C2, and C3 vertebral levels in healthy controls (CTRL, blue) and multiple sclerosis (MS, orange) cohorts. Negative SRM values indicate longitudinal reduction in CSA over time, whereas positive values indicate stability or increase. Error bars represent 95% confidence intervals of the SRM estimates.

error classification model solution requires further optimization to achieve full automation, the current pipeline highlights the most questionable outputs for visual review, enhancing the efficiency of manual inspection. This is particularly important in multi-site clinical studies or population-based cohorts, where the volume of data necessitates scalable QC solutions[23]. Additionally, the computation of image quality metrics specifically optimized for the SC region allows users to better identify inconsistencies or degradations that could affect the accuracy of downstream analyses. Threshold values were established to serve as reference points for categorizing image quality as either good or low-quality. In many cases, image artifacts, noise, or motion-related distortions may be present outside the cord area, but have minimal to no effect on SC segmentation and labeling. By focusing the quality assessment on the region of interest, the pipeline ensures that researchers can prioritize meaningful quality issues without discarding MRI volumes unnecessarily. This approach enhances efficiency and promotes more informed decisions about data inclusion, which is especially critical in studies with limited or heterogeneous datasets. Importantly, the accuracy of the VGG-based label error classification model is less in low-quality MRI volumes, increasing the likelihood of false outputs. This observation highlights the importance of using the VGG-based label error classification model in tandem with image quality metrics to select the subset of MRI volumes requiring manual review (i.e., all MRI volumes graded as low-quality should be reviewed or excluded).

The availability of an objective, automated approach for quantitative assessment of SC is critical for translation of outcomes to clinical care and trial settings. As a proof of principle, the developed tool showed sensitivity to longitudinal changes in CSA in MS individuals, replicating previous studies in the same population[5,24,25]. This proof-of-concept analysis in MS was exploratory and did not adjust covariates, but provides preliminary evidence of applicability in a real-world effort. Similar use-cases in other neuromotor-related diseases, such as ALS and inherited cerebellar ataxias, are compelling given the primary involvement of SC degeneration in the pathology of these conditions. For example, a recent phase II clinical trial in individuals with Friedreich ataxia nominated SC CSA as the primary outcome measure[22]. The continued standardization and validation of imaging endpoints for use in clinical trials remains a pressing need and active area of investigation in neurodegenerative diseases[26]. The availability of this pipeline will aid in addressing this need in diseases that impact the spinal cord. From a practical standpoint, CPU-only processing results in longer computation times compared to GPU-accelerated execution; this is unlikely to substantially limit real-world use, as most applications do not require immediate results. Furthermore, cloud-based deployment may provide a viable alternative for sites without access to dedicated GPU resources. However, such approaches require careful consideration of legal and regulatory frameworks related to data protection, patient privacy, and secure data transfer.

The main limitation of this study is the lack of a ground truth dataset, especially for C1 vertebrae. Although there is no consensus protocol to identify the C1 level, we used a well-curated and diverse dataset, similar to previous successful tools[27]. However, ground truths are required to further future improvements. Inter-rater reliability for vertebral labeling could not be formally quantified, as pairwise agreement records between raters were not retained, preventing retrospective computation of agreement metrics (e.g., Cohen's κ). Although a conservative consensus strategy was used to minimize false-positive labels, this limitation may affect the assessment of annotation consistency, particularly for anatomically ambiguous regions such as the C1 level. In addition, the observed correlation between the nnU-Net model and SCT reflects performance in cases where both methods successfully generated valid outputs and may therefore overestimate agreement in more challenging scenarios. The VGG-based label error classification model currently also has only moderate specificity, and further developments will be necessary to reduce the false negative rate. Novel DL architectures may be revisited as larger imaging datasets become available. The proposed QC thresholds ($SNR < 48$ and $EFC > 8.35 \times 10^{-5}$) were derived from the internal dataset and may require further validation to confirm their generalizability across external datasets and acquisition protocols. Additionally, internal data splitting for model training and validation was performed at the subject level, without site- or protocol-aware partitioning. While models trained on multi-site data can sometimes be influenced by acquisition site- or protocol-related dependencies, our model's robustness and generalizability are supported by comprehensive out-of-sample testing on five independent external datasets acquired at multiple sites with heterogeneous imaging protocols. Another limitation is the inclusion of exclusively diagnosed inherited cerebellar ataxias and MS in the disease group. Generalizability to other pathologies may require further validation, especially traumatic SC injuries. Future work exploring integration with other imaging modalities (e.g., diffusion-weighted imaging) may help further enhance clinical utility.

In conclusion, our work provides state-of-the-art performance in automated C1–C3 vertebral labeling and spinal cord morphometric extraction from in vivo MRI of the human head. This framework complements existing SC segmentation tools by improving vertebral level localization and enabling accurate downstream quantification. This outcome offers a valuable tool for characterizing SC neuroanatomy in health and disease and can be readily deployed and scaled for use in multi-site and large population datasets. Our solution provides a pathway to support the validation of SC imaging biomarkers for monitoring disease progression and evaluating treatment efficacy in neurological disorders.

Methods

Data for training, internal validation, and testing

3D T1-weighted MRI scans of the whole head were obtained from three different international consortia: ENIGMA-Ataxia[3,4], TRACK-FA[28] and READISCA[15]. Participant recruitment and the original inclusion and exclusion criteria were defined by each source study and are described in the respective publications[3,4,15,28]. No additional participant eligibility criteria were applied after dataset aggregation. These images were acquired at 32 centers using 8 different MRI protocols on Siemens or Philips scanners (Table S11). The collected dataset included 1877 participants: 998 healthy controls and 879 subjects with molecular confirmation of an inherited cerebellar ataxia, including 354 FRDA and different spinocerebellar ataxias (SCA) (119 SCA1, 87 SCA2, 246 SCA3, 25 SCA6, and 48 SCA7 patients) (Table 1). The 1877 individuals contributed exactly 1877 3D MRI volumes, corresponding to one volume per subject. All MRI volumes were de-identified and in NIfTI format with no embedded personal health information. The diversity of inherited cerebellar ataxias included in the dataset provides anatomical variability, allowing for broad generalizability of the developed algorithm. Of the total 1877 MRI volumes, 288 were acquired

with an isotropic voxel spacing of 0.8 mm, while the remaining 1,589 MRI volumes had a voxel spacing of 1.0 mm isotropic.

The study was conducted in accordance with the Declaration of Helsinki, and it was approved by the Ethics Committee of the School of Medicine at the University of São Paulo at Ribeirão Preto (registered CAAE 92205925.6.0000.5440). The ethics committee or Institutional Review Board (IRB) respective to each project data source or site approved the collection and use of data for research purposes. Written informed consent to participate in the study was obtained from all participants at each site, or from a parent or legal guardian for participants under 16 years of age, prior to original data collection.

The READISCA[15–17] steering committee approved the data use, and IRB approval was obtained from the Institutional Review Board of the University of Minnesota (IRB protocol number: STUDY00002420, Clinical Trial Readiness for SCA1 and SCA3, approval date: 2/15/2018).

The TRACK-FA steering committee approved the data use, and IRB reference numbers were previously published[9] (Monash Health Human Research Ethics Committee: RES-20-0000-139A; Children's Hospital of Philadelphia: IRB 20-017611; University of Minnesota: IRB STUDY00009047; University of Florida: IRB202000399; RWTH Aachen University: EK195/20; University of Campinas: CAAE NO 83241318.3.1001.5404; McGill University: 2022-8676).

Ethics approvals from ENIGMA-Ataxia sites contributing to this study are: Campinas: CAAE 83241318.3.1001.5404; Melbourne: Monash Health HREC 13201B, Monash University HREC 7810 and 12372; University of Minnesota: 1210M22281; Essen: EC Medical Faculty University of Duisburg-Essen 15-6404-BO; Minnesota: 0502M67488; Milan: IRB approval 17 December 2014 (N. 14); Curitiba: CAAE 06128812.0.0000.0096; Paris: local ethical committee AOM10094, CPP Ile de France VI, Ref: 105–10; Nijmegen: CMO regio Arnhem-Nijmegen, file number 2010/438; Tubingen: IRB University Hospital Tübingen, 303/2008BO2. Participants gave informed consent to participate in the study before taking part.

Spinal cord vertebral labeling

The MRI volumes were initially preprocessed using the SCT version 6.2[2] to obtain the SC segmentation and vertebral label identification masks. However, visual inspection revealed notable errors (identified by GMJ and TJRR), with 47% of the masks from ataxic individuals (417 of 879) and 28% from healthy controls (281 of 998) having mislabeled vertebral levels (Figs. 5 and S7). These 698 MRI volumes, representing a deliberately selected subset of challenging cases that failed standard automated processing, were used as the test dataset to evaluate the robustness of the deep learning models under complex imaging and anatomical conditions.

A set of 924 accurately labeled MRI volumes by SCT (GMJ, all 462 patients and randomly selected 462 controls, matched to the patient group to ensure class balance and reduce the risk of model overfitting during training) formed the training and validation dataset (101 with 0.8 mm and 823 with 1.0 mm isotropic voxel resolutions), including representation from all cohorts: 462 healthy individuals, 177 FRDA, 62 SCA1, 49 SCA2, 139 SCA3, 10 SCA6, and 25 SCA7. The labeled masks from SCT were refined to retain only C1 to C3 levels. This adjustment was necessary to account for the lack of C4 coverage in some MRI volumes, which could lead to sampling issues and potentially impact the model's performance. The dataset was split at the subject level, as each participant contributed a single 3D image. Subjects were randomly shuffled and assigned to one of five folds for cross-validation. Approximately 80% of the subjects ($n \approx 739$) were used for training and 20% ($n \approx 185$) for validation in each fold.

Vertebral level landmarks and identification

Most of the studies assessing upper spinal cord use vertebral levels as an anatomical reference[29,30]. Although useful, this approach has been shown to introduce variability in spinal cord-derived metrics, especially cross-sectional area[10,31]. New strategies have been proposed to attempt to ameliorate this variability, such as using the distance from the pontomedullary junction[32]. Despite the improvement, C1 level remained neglected from the analysis and still works on a same standard distance for all subjects, with low sensitivity to intrinsic anatomical variations[10,32]. Since C1 lacks consistent standardization due to the absence of a well-defined intervertebral disc[10,32], we proposed a standardized vertebral level localization protocol based on anatomical landmarks visible in sagittal T1-weighted MRI. It is important to say we did not perform any ground-truth dataset, but a curated data training set, similar to what FastSurfer developers did[27]. For C1, the superior border was considered a horizontal line on sagittal images as reference the occipito-atlantal joint in the posterior part of the foramen magnum. For the inferior border, we considered another horizontal line on sagittal images down to the atlanto-axial joint line across the dens in the C1 arch (Fig. S7). For C2, the superior border follows the atlanto-axial joint line in the C1 arch through the C2–C3 intervertebral disc (Fig. S7). Lastly, the C3 level follows the inferior end of the cervical C2-C3 disc to the next lower cervical disc (C3-C4).

Masks were independently assessed by two experienced neuroimaging scientists (GMJ and DHS) and were considered incorrect if any of the three labeled vertebral levels (C1–C3) were clearly misplaced relative to their expected anatomical positions (i.e., positioned too cranially, too caudally, or incompletely labeled) when evaluated in the sagittal plane. All raters were blinded to diagnosis. A conservative voting strategy was applied to assign final labels: a mask was classified as correct only if both raters independently agreed on its correctness, whereas any disagreement or assignment of *incorrect* by either rater resulted in a final classification of incorrect. This curated "silver-standard" annotation strategy is consistent with established practices in spinal cord imaging pipelines and has been adopted in prior studies using similar validation frameworks[3,4,9,33]. Most labeling masks generated by SCT showed substantial variability for the C1 vertebral level, frequently placing it either too high (including part of the brainstem) or too low relative to its probable location. This misplacement affects the accuracy of downstream analyses, particularly the computation of spinal cord cross-sectional area. In cases where the C1 level includes brainstem tissue, the axial area measurements for this level become artificially inflated, thus introducing significant bias in the mean area calculation and compromising statistical analyses. Fig. S8 illustrates the anatomical references used to classify each labeling mask as correct or incorrect. Our analyses focus exclusively on the first three cervical levels (C1–C3).

Automated quality control

The dataset used for the automated quality control DL model included 1396 MRI volumes, comprising 698 randomly selected successful cases and all 698 failed cases (Figs. S7 and S8). The dataset was divided randomly at the subject level into training (70%, $n = 977$), validation (20%, $n = 279$), and test (10%, $n = 140$) sets with no possibility of cross-contamination (Fig. 5).

Quantitative metrics describing the image quality in the SC region were computed using well-established image quality metrics typically applied to brain MRI[23], including: SNR, contrast-to-noise ratio (CNR), CJV, and EFC. These metrics were calculated within a region of interest (ROI) defined around the SC, with the derived label mask serving as a centroid reference (see supplementary material for detailed information). The MRIs were visually inspected and classified as high, medium, and low quality (GMJ and DHS) (Fig. S9). Next, 30 MRI volumes were randomly selected as good-, medium-, or low-quality to assess the utility of such metrics (see "Statistical analysis," below). This analysis was repeated five times to establish the robustness of the imaging quality metrics.

Image quality metrics

The SNR quantifies the clarity or quality of a signal in an image relative to the level of background noise. In MRI, SNR specifically quantifies how distinguishable the tissue signal is from the random fluctuations or noise present in the image[34]. A higher SNR indicates a clearer and more reliable signal, which is crucial for accurate image interpretation and analysis; in this pipeline we calculated the SNR Dietrich method[34] which includes a correction factor in the estimation of the standard deviation of air and its

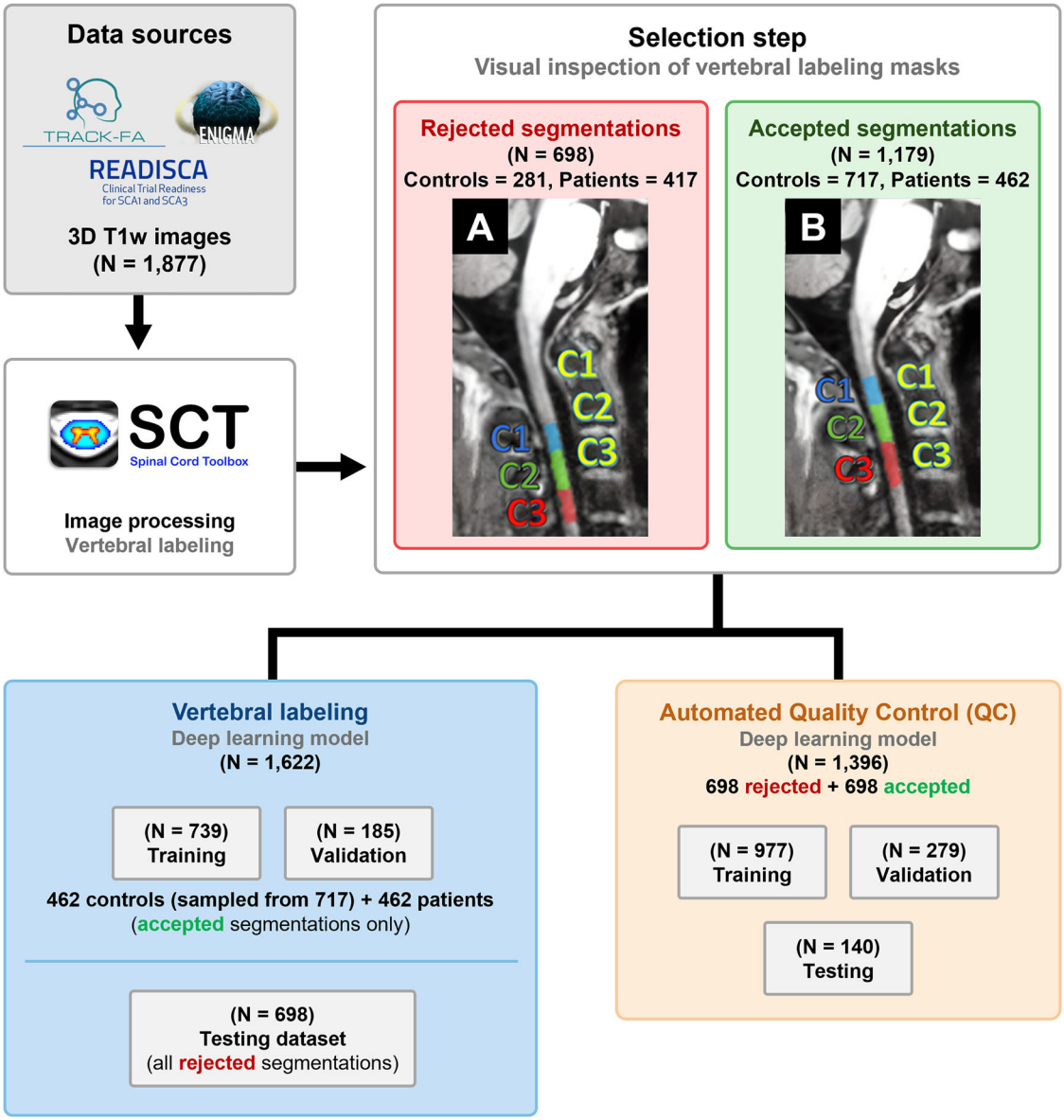


Fig. 5 | Data workflow. A total of 1877 T1-weighted MRI volumes were processed using the Spinal Cord Toolbox (SCT), followed by visual inspection to separate the data. Of the 1,179 MRI volumes accepted after visual quality control, 924 were selected for a balanced labeling training dataset (462 controls and 462 patients). The 698 rejected MRI volumes formed the testing dataset. The same approach was used for training the automated quality control (QC) model (1396 MRI volumes divided equally between 698 rejected and 698 accepted — randomly subsampled from the pool of 1,179 accepted masks — labeled volumes). The presented images show examples of misidentified (**A**) and correctly identified (**B**) vertebral labels (C1, C2, and C3) in an individual with Friedreich's Ataxia (FRDA). *N* number of images.

Rayleigh distribution:

$$SNR = \frac{\mu_F}{\sqrt{\frac{2}{4-\pi}}\sigma_{air}}$$

where μ_F is the mean intensity of the foreground, σ_{air} is the standard deviation of the air surrounding the head ("hat" mask).

Background air and foreground regions used for SNR estimation were defined using a fully automated, intensity-based procedure implemented in Python with the SciPy 1.15.0 library for morphological image processing. All MRI volumes were first intensity-normalized by dividing voxel intensities by the maximum intensity value in the image. An initial head mask was generated by thresholding voxels with normalized intensity values greater than 0.05, based on the assumption that background air exhibits near-zero signal intensity in T1-weighted MRI. To account for partial-volume effects and uncertainty at tissue boundaries, the head mask was iteratively expanded using binary dilation. A second, larger dilation was then applied to define an outer region surrounding the head. The background ("hat") mask was constructed as the shell between the inner and outer dilated regions, with all voxels belonging to the original head mask explicitly excluded. This procedure yields a hollow ring around the head containing only background air voxels, which was used to estimate background noise.

The foreground region was defined using the same normalized intensity threshold (>0.05), classifying all voxels with non-zero anatomical signal as foreground. This conservative delineation of the entire head region is robust across subjects, scanners, and acquisition protocols and was used to compute the mean foreground signal required for SNR calculation.

The CNR measures the contrast between the cerebrospinal fluid (CSF) and SC in relation to the overall noise level in the image, indicating how well

these regions can be differentiated, calculated as follows[35]:

$$CNR = \frac{|\mu_{CSF} - \mu_{SC}|}{\sqrt{\sigma_B^2 + \sigma_{SC}^2 + \sigma_{CSF}^2}}$$

where μ_{CSF} is the mean intensity of the CSF, μ_{SC} is the mean intensity of the spinal cord, σ_B^2 is the standard deviation of the background noise, σ_{SC}^2 is the standard deviation of the spinal cord intensity, σ_{CSF}^2 is the standard deviation of the CSF intensity. In this study, we applied this CNR formulation to an optimized region of the spinal cord, specifically chosen to maximize contrast and minimize confounding effects such as partial volume artifacts. This anatomical approximation was adopted to reflect the local contrast between the spinal cord and its immediate environment, providing a robust and consistent measure of image quality across subjects.

The CJV quantifies the degree of overlap between the intensity distributions of CSF and spinal cord[36]. A higher value indicates more overlap, suggesting less distinction between these two regions, which can affect image quality and the accuracy of segmentation tasks, given by[36]:

$$CJV = \frac{\sigma_{SC} + \sigma_{CSF}}{|\mu_{SC} - \mu_{CSF}|}$$

where σ_{SC} is the standard deviation of the spinal cord intensity, σ_{CSF} is the standard deviation of the CSF intensity, μ_{SC} is the mean intensity of the spinal cord, μ_{CSF} is the mean intensity of the CSF. As with the CNR calculation, we applied this measure to an optimized region of the spinal cord. This anatomical proxy allows for consistent and meaningful estimation of contrast-based metrics in the spinal cord.

The EFC relies on the Shannon entropy[37] of voxel intensities to evaluate the impact of ghosting and blurring caused by head movement. Lower EFC values are preferable, with EFC = 0 indicating that all the image's energy is concentrated in a single voxel, signifying an ideal focus without motion artifacts, calculated as follows:

$$E = -\sum_{j=1}^{N} \frac{x_j}{x_{\max}} \ln\left[\frac{x_j}{x_{\max}}\right]$$

where E is the Shannon entropy of the voxel intensities in the image. It provides a measure of the intensity distribution of the voxels, with lower values indicating greater concentration of intensity in a few voxels and higher values indicating a more uniform distribution, where:

$$x_{\max} = \sqrt{\sum_{j=1}^{N} x_j^2}$$

x_j is the intensity value of voxel j in the image, $x_{\max}$ is the maximum possible intensity value in the image, used to normalize the voxel intensities, the logarithmic term $\ln\left[\frac{x_j}{x_{\max}}\right]$ captures the entropy, where higher values indicate greater disorder or blurring. The equation is scaled by the maximum possible entropy, allowing the EFC to be compared consistently across images of varying sizes and dimensions:

$$EFC = \left(\frac{N}{\sqrt{N}} \ln \sqrt{N}^{-1}\right) E$$

where N refers to the total number of voxels in the image, $\sqrt{N}$ is used to adjust the scale of entropy in the formula and $\ln\sqrt{N}^{-1}$ allows for comparing entropy between MRI volumes of different dimensions.

To evaluate the utility of the proposed methods, we visually inspected and classified all the MRI volumes as high-, medium-, and low-quality (GMJ and DHS) (Fig. S9). High-quality MRI volumes are characterized by the clear absence of noise, graininess, and subject motion, with sharp contrast allowing unambiguous distinction of all anatomical structures. Medium-quality MRI volumes exhibit partial motion effects and reduced contrast, particularly in the intervertebral disc regions. Low-quality MRI volumes show a complete loss of anatomical reference in the spinal cord region, caused by noise, poor contrast, and motion artifacts. Next, 30 MRI volumes were randomly selected as good-, medium-, or low-quality to assess the utility of such metrics. This analysis was repeated five times to establish the robustness of the imaging quality metrics.

Deep learning architectures for the vertebral labeling model

All three 3D segmentation models, nnU-Net v2[12], SwinUNETR[38], and a 3D U-Net[11] with deep supervision (DS)[39], were trained using an identical preprocessing pipeline to ensure a fair and controlled comparison across architectures. The input data consisted of T1w brain MRI volumes acquired at isotropic spatial resolutions of 0.8 mm and 1.0 mm, together with corresponding 3D spinal cord vertebral labeling masks encompassing the first three cervical vertebral levels (C1–C3).

To harmonize spatial resolution across acquisitions, all MRI volumes were resampled to a common isotropic voxel spacing of 1.0 mm. This choice was driven by the distribution of the data, as the majority of scans were acquired at 1.0 mm resolution, while only a smaller subset was available at 0.8 mm. Resampling the minority resolution was therefore preferred over downsampling the entire dataset. Resampling was performed using third-order spline interpolation for image volumes and nearest-neighbor interpolation for segmentation masks to preserve label integrity. Following resampling, all volumes were reoriented to a canonical orientation. Intensity normalization was applied using z-score normalization, computed on a per-volume basis within the non-zero foreground region, to reduce inter-subject intensity variability while preserving tissue contrast.

The architecture experimentation task involved exploring various models, including U-Net[11] with DS[39], SwinUNETR[38], and the nnU-Net framework[12], all 3D architectures. The U-Net with DS architecture began with feature channels of 32, doubling progressively to 512. Each layer consisted of two $3 \times 3 \times 3$ convolution operations, followed by instance normalization[40] and activation via a Leaky ReLU[41] with a 0.01 negative slope. Training employed a Stochastic Gradient Descent optimizer with a learning rate of 3×10^{-2} and a Nesterov momentum[42,43] of 0.99, alongside the Dice loss function. The learning rate was reduced using a polynomial decay scheduler. The model was trained for over 600 epochs with a batch size of 1. For the SwinUNETR, the original implementation was adjusted to a feature size of 24. The AdamW[44] optimizer was applied with a learning rate of 1×10^{-4}, weight decay of 1×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. The loss function used was Dice loss, and a cosine annealing scheduler was implemented for the learning rate. This model underwent training for 1000 epochs with a batch size of 1. The nnU-Net framework was trained using its default settings for the 3D full-resolution trainer. For nnU-Net, patch size and batch size were automatically optimized based on available GPU memory, following the framework's default configuration. For SwinUNETR and the 3D U-Net with deep supervision, the same patch size was explicitly adopted to ensure architectural consistency across models. No additional resizing of the full volumes was performed beyond resampling and patch-based cropping. For both U-Net with DS and SwinUNETR, the same augmentation techniques were applied, including random rotation, shearing, scaling, translation, flipping, elastic deformation, and gamma contrast, replicating the augmentation pipeline used by nnU-Net.

Deep learning architectures for the automated quality control model

All experiments were conducted using single 2D mid-sagittal slices derived from T1w spinal cord MRI. Image–mask pairs were generated using the quality control module of the SCT, which outputs a single 2D PNG mid-sagittal slice with voxel intensities contrast-rescaled for visualization. During QC generation, SCT applies a linear per-slice rescaling of voxel intensities to an 8-bit grayscale range (0–255), based on the minimum and maximum intensity values within each slice. While this operation standardizes the dynamic range for visualization, it does not represent an intensity

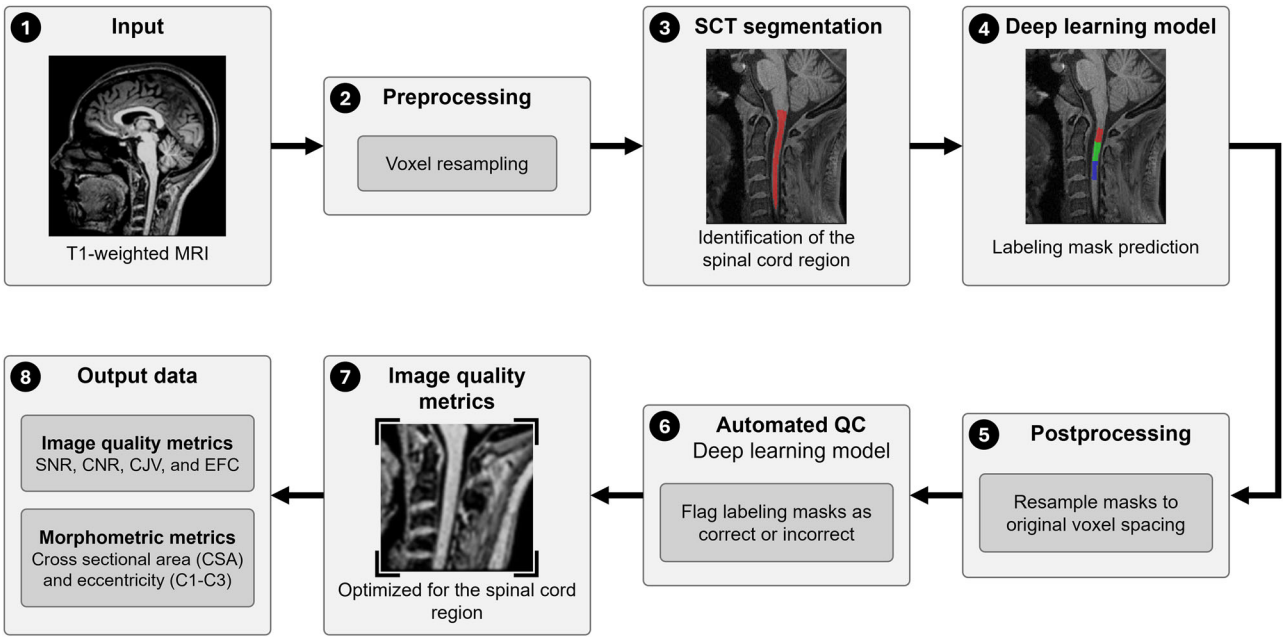


Fig. 6 | Schematic workflow of the image processing pipeline. Representative schematic of the image processing pipeline. The input MRI volumes are resampled to 1.0 mm isotropic if necessary, followed by *z*-score intensity normalization and segmentation of the spinal cord region using the Spinal Cord Toolbox (SCT) version 6.2. The deep learning prediction model then generates a labeled mask for the first three vertebral levels (C1–C3). This is followed by resampling back to the original MRI volume resolution if necessary. The deep learning automated quality control (QC) step is then undertaken, which flags the labeling mask as correct or incorrect. The final step computes image quality metrics (SNR, CNR, CJV, and EFC) in the spinal cord region and outputs the cross-sectional area and eccentricity values for each vertebral level. SNR signal-to-noise ratio, CNR contrast-to-noise ratio, CJV coefficient of joint variation, EFC entropy focus criterion.

normalization procedure suitable for machine learning. Therefore, prior to model training, all single mid-sagittal slices underwent an explicit *z*-score normalization step to ensure numerical stability and consistency across samples.

An automated preprocessing pipeline was applied to each image–mask pair. First, a slice-wise bounding box was computed directly from the spinal cord segmentation mask by identifying the minimal enclosing region containing all non-zero voxels. This bounding box was symmetrically expanded by a fixed margin to ensure full coverage of the spinal cord and surrounding anatomical context. To maintain spatial consistency across subjects and slices, bounding box dimensions were constrained to predefined minimum and maximum limits. Bounding boxes exceeding these limits were clipped accordingly, whereas smaller boxes were padded to reach the target size. This strategy ensured complete spinal cord inclusion while reducing background variability and standardizing the FoV prior to resizing.

Cropped single mid-sagittal slices and corresponding masks were resized to 224 × 224 pixels. Bilinear interpolation was applied to single mid-sagittal slices, while nearest-neighbor interpolation was used for masks to preserve discrete class labels. The choice of this input resolution was motivated by the architectural requirements of CNNs initialized with ImageNet pre-trained weights, particularly VGG-based models, which require fixed input dimensions to enable direct reuse of pre-trained convolutional filters.

Segmentation masks were converted to a four-channel one-hot encoding (background plus three anatomical classes) and concatenated with the image channel, resulting in a five-channel input tensor. Data augmentation was applied on-the-fly during training using the Albumentations library (2.0.8) and included random horizontal flips, small rotations (±10°), Gaussian noise, motion blur, and random brightness and contrast adjustments. For training, three 2D CNNs were explored: EfficientNet-B7[45], ResNet-50[46], and VGG-16[47]. To construct the dataset, a single mid-sagittal slice was extracted from each participant's 3D T1w MRI volume and saved as a single 2D grayscale PNG mid-sagittal slice. Corresponding vertebral labeling masks (C1–C3) were processed to generate aligned single mid-sagittal 2D PNG segmentation masks, which were one-hot encoded and concatenated with the single grayscale mid-sagittal slice, resulting in 5-channel inputs. For EfficientNet-B7 and VGG-16, the first convolutional layer was modified to accept 5 channels, and the classifier was replaced with a dropout layer ($p = 0.5$) followed by a linear layer for binary classification. The ResNet-50 model retained its original feature extractor, with only the final fully connected layer replaced. All models were fine-tuned using the Adam optimizer (learning rate = 1×10^{-4}, weight decay = 1×10^{-5}), a batch size of 32, and a weighted cross-entropy loss to address class imbalance. A ReduceLROnPlateau learning rate scheduler and early stopping strategy (patience = 100 epochs) were employed[48]. Single mid-sagittal slices were augmented with horizontal flipping, random rotation, Gaussian noise, motion blur, and brightness/contrast adjustments.

Additionally, for the labeling step, it is only relevant to determine whether the vertebral level is correctly identified in the sagittal view, without considering whether the area is fully filled in the axial view. This is because SCT uses the binary segmentation mask of the entire spinal cord, which is generated separately, to perform subsequent registration to a template. Therefore, the labeling step solely requires the correct identification of vertebral levels, and not the precise delineation of their axial cross-sectional areas.

Pipeline strategy

The pipeline workflow (Fig. 6) begins with preprocessing of the input brain 3D T1w MRI, including resampling to 1.0 mm isotropic image resolution, intensity z-normalization, and SC identification using the SCT version 6.2[2]. Next, a multiclass vertebral labeling mask prediction is performed using the trained deep learning model (C1–C3). After the prediction, both the output mask and the input MRI volume are resampled back to the original voxel resolution. The automated QC step is then performed to check the vertebral labeling identification as either correct or incorrect, along with the derivation of image quality metrics (SNR, CNR, CJV, EFC). Finally, the cross-sectional area (CSA) and eccentricity values for C1–C3 are extracted. The workflow uses Python 3.11.2, SciPy 1.15.0, NiBabel 5.2.1, and PyTorch 2.6.0.

Data for external testing

To rigorously evaluate the generalizability of the vertebral labeling and automated QC models, we gathered 2737 additional brain 3D T1w MRIs from public repositories, including the C-MIND (Cincinnati MR Imaging of NeuroDevelopment)[18], IXI (Information eXtraction from Images) (https://brain-development.org/ixi-dataset/), HCP-D (Human Connectome Project-Development)[49,50], and the NIH (National Institutes of Health (NIH) MRI study of normal brain development)[19] datasets (Tables S1 and S9). We also evaluated 51 cases of Multiple Sclerosis (MS) not included in the training datasets from the University of Campinas to demonstrate reproducibility and generalizability across disease conditions (Tables S1 and S9). Each validation dataset used in this study presents distinct characteristics and acquisition parameters, which is ideal for testing the reproducibility and robustness of our model compared to SCT (Table S9 and Fig. S6). The C-MIND dataset includes MRI volumes in which the lower portion of the SC is cropped, allowing visualization in most cases only of the C1 vertebral level and rarely C2. The NIH dataset contains lower-quality MRI volumes acquired on 1.5 T scanners, representing a more challenging acquisition setting. The HCP dataset is the largest and exhibits the highest inter-individual variability, making it particularly useful for evaluating model generalization across diverse populations. The IXI dataset provides low-resolution MRI volumes with anisotropic voxel spacing, which significantly differs from the characteristics present in the training dataset. Additionally, patients with MS present distinct SC characteristics compared to the ataxic cohort used for training, such as focal demyelinating lesions and variable inflammation patterns, rather than the diffuse SC atrophy typically observed in hereditary ataxias.

All outputs generated (from the proposed model and/or SCT) on the external datasets were visually inspected by two experienced neuroimaging scientists (GMJ and DHS) and classified as correct or incorrect. A result was considered *correct* when vertebral labels for C1–C3 were correctly segmented and anatomically aligned, and *incorrect* when labels were misplaced, partially missing, or anatomically inconsistent. Annotation followed the proposed standardized vertebral level localization protocol, based on explicit anatomical landmarks visible on sagittal T1-weighted MRI, focusing exclusively on the first three cervical levels (C1–C3) (see supplementary material for detailed information). This curated "silver-standard" evaluation strategy is consistent with prior spinal cord imaging pipelines and large-scale validation studies[3,4,9,33].

Statistical analysis

To evaluate the performance of our deep learning models for SC vertebral labeling, we utilized metrics commonly used in medical image segmentation, including the Dice Similarity Coefficient[51,52], Intersection over Union[53], precision, recall, and Hausdorff distance[54], using as reference the curated high-quality vertebral masks. To assess the automated QC pipeline, we quantified accuracy (ratio of correctly classified true predictions to the total number of predictions), specificity, and sensitivity[13]. ROC curve analysis was used to assess the performance of each QC metric in distinguishing low-quality MRI volumes from acceptable cases, using expert-curated labels as the reference standard. For each metric, sensitivity and specificity were computed across a range of threshold values. Optimal cut-off points for low-quality classification were selected using the Youden index, maximizing the sum of sensitivity and specificity. MRI volumes with metric values exceeding (or falling below, depending on the directionality of the metric) the selected threshold were classified as low quality and excluded from further analysis. To compare the performance metrics across the different algorithms, we used a paired *t*-test. Pearson correlation coefficients were employed to evaluate associations between SC morphometric data. Bonferroni correction was applied to account for multiple comparisons, with corrected *p*-values < 0.05 considered statistically significant.

To evaluate the longitudinal sensitivity of our final model, longitudinal slopes of CSA and eccentricity were calculated using per-subject linear regression (one slope per subject and per metric, using all available time points for that subject). Within-group and between-group differences in slopes were assessed using two-sided one-sample and two-sample *t*-tests, respectively. Longitudinal effect sizes were obtained using standardized response means (SRM = mean slope/SD slope) for within-group changes and Cohen's *d* for between-group differences in slopes. No correction for multiple comparisons was used for this exploratory analysis.

Statistical analysis was conducted using Python 3.11.5.

Data availability

Data generated or analyzed during the study are available from the corresponding author by request. De-identified participant-level clinical and imaging data from READISCA and a data dictionary defining each field in the set are available from the National Institute of Mental Health Data Archive (https://nda.nih.gov/,Collection C3155). TRACK-FA (NGK) and ENIGMA-Ataxia (IHH) data analyzed during the study were provided by a third party.

Code availability

This tool is publicly available at https://github.com/art2mri/Enigma-SC.

Received: 24 November 2025; Accepted: 1 October 2026;
Published online: 23 October 2026

References

1. Kandel, E. R. *Principles of Neural Science* 5th edn (McGraw-Hill Medical, 2013).
2. De Leener, B. et al. SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data. *Neuroimage* **145**, 24–43 https://doi.org/10.1016/j.neuroimage.2016.10.009 (2017).
3. Rezende, T. J. R. et al. Genotype-specific spinal cord damage in spinocerebellar ataxias: an ENIGMA-Ataxia study. *J. Neurol. Neurosurg. Psychiatry* **95**, 682–690 https://doi.org/10.1136/jnnp-2023-332696 (2024).
4. Rezende, T. J. R. et al. Progressive spinal cord degeneration in Friedreich's Ataxia: results from ENIGMA-Ataxia. *Mov. Disord.* **38**, 45–56 https://doi.org/10.1002/mds.29261 (2023).
5. Taheri, K. et al. Cervical spinal cord atrophy can be accurately quantified using head images. *Mult. Scler. J. Exp. Transl. Clin.* **8**, 20552173211070760 https://doi.org/10.1177/20552173211070760 (2022).
6. Wimmer, T. et al. The upper cervical spinal cord in ALS assessed by cross-sectional and longitudinal 3T MRI. *Sci. Rep.* **10**, 1783 https://doi.org/10.1038/s41598-020-58687-z (2020).
7. Jovicich, J. et al. Brain morphometry reproducibility in multi-center 3T MRI studies: a comparison of cross-sectional and longitudinal segmentations. *Neuroimage* **83**, 472–484 https://doi.org/10.1016/j.neuroimage.2013.05.007 (2013).
8. Elliott, L. T. et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216 https://doi.org/10.1038/s41586-018-0571-7 (2018).
9. Georgiou-Karistianis, N. et al. Neuroimaging biomarkers for Friedreich Ataxia: a cross-sectional analysis of the TRACK-FA Study. *Ann. Neurol.* **98**, 386–397 https://doi.org/10.1002/ana.27237 (2025).
10. Ullmann, E., Pelletier Paquette, J. F., Thong, W. E. & Cohen-Adad, J. Automatic labeling of vertebral levels using a robust template-based approach. *Int. J. Biomed. Imaging* **2014**, 719520 https://doi.org/10.1155/2014/719520 (2014).
11. Ronneberger, O., Fischer, P. & Brox, T. in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* 234–241 (Springer).
12. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 https://doi.org/10.1038/s41592-020-01008-z (2021).

13. Petrov, D. et al. in *International Workshop on Shape in Medical Imaging* 268–276 (Springer).
14. Food and Drug Administration. *Clinical Trial Imaging Endpoint Process Standards Guidance for Industry*. https://www.fda.gov/media/81172/download.
15. Chandrasekaran, J. et al. Clinically meaningful magnetic resonance endpoints sensitive to Preataxic Spinocerebellar Ataxia Types 1 and 3. *Ann. Neurol.* **93**, 686–701 https://doi.org/10.1002/ana.26573 (2023).
16. Rezende, T. J. R. et al. Sensitivity of advanced magnetic resonance imaging to progression over six months in early Spinocerebellar Ataxia. *Mov. Disord.* **39**, 1856–1867 https://doi.org/10.1002/mds.29934 (2024).
17. Rezende, T. J. R. et al. Value of MRI outcomes for preventive and early-stage trials in Spinocerebellar Ataxias 1 and 3. *Ann. Clin. Transl. Neurol.* https://doi.org/10.1002/acn3.70325 (2026).
18. Vannest, J. et al. Factors determining success of awake and asleep magnetic resonance imaging scans in nonsedated children. *Neuropediatrics* **45**, 370–377 https://doi.org/10.1055/s-0034-1387816 (2014).
19. Evans, A. C. & Brain Development Cooperative, G. The NIH MRI study of normal brain development. *Neuroimage* **30**, 184–202 https://doi.org/10.1016/j.neuroimage.2005.09.068 (2006).
20. Forsberg, D., Sjoblom, E. & Sunshine, J. L. Detection and labeling of vertebrae in MR images using deep learning with clinical annotations as training data. *J. Digit. Imaging* **30**, 406–412 https://doi.org/10.1007/s10278-017-9945-x (2017).
21. Valosek, J., Bedard, S., Kerkovsky, M., Rohan, T. & Cohen-Adad, J. A database of the healthy human spinal cord morphometry in the PAM50 template space. *Imaging Neurosci.* **2**, https://doi.org/10.1162/imag_a_00075 (2024)
22. Pandolfo, M. et al. Efficacy and safety of leriglitazone in patients with Friedreich Ataxia: a phase 2 double-blind, randomized controlled trial (FRAMES). *Neurol. Genet.* **8**, e200034 https://doi.org/10.1212/NXG.0000000000200034 (2022).
23. Esteban, O. et al. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS ONE* **12**, e0184661 https://doi.org/10.1371/journal.pone.0184661 (2017).
24. Oh, J. et al. Five-year longitudinal changes in quantitative spinal cord MRI in multiple sclerosis. *Mult. Scler.* **27**, 549–558 https://doi.org/10.1177/1352458520923970 (2021).
25. Valsasina, P., Rocca, M. A., Horsfield, M. A., Copetti, M. & Filippi, M. A longitudinal MRI study of cervical cord atrophy in multiple sclerosis. *J. Neurol.* **262**, 1622–1628 https://doi.org/10.1007/s00415-015-7754-z (2015).
26. Oz, G. et al. MR imaging in Ataxias: consensus recommendations by the Ataxia Global Initiative Working Group on MRI Biomarkers. *Cerebellum* **23**, 931–945 https://doi.org/10.1007/s12311-023-01572-y (2024).
27. Henschel, L. et al. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. *Neuroimage* **219**, 117012 https://doi.org/10.1016/j.neuroimage.2020.117012 (2020).
28. Georgiou-Karistianis, N. et al. A natural history study to track brain and spinal cord changes in individuals with Friedreich's ataxia: TRACK-FA study protocol. *PLoS ONE* **17**, e0269649 https://doi.org/10.1371/journal.pone.0269649 (2022).
29. Casserly, C. et al. Spinal cord atrophy in multiple sclerosis: a systematic review and meta-analysis. *J. Neuroimaging* **28**, 556–586 https://doi.org/10.1111/jon.12553 (2018).
30. Moccia, M. et al. Advances in spinal cord imaging in multiple sclerosis. *Ther. Adv. Neurol. Disord.* **12**, 1756286419840593 https://doi.org/10.1177/1756286419840593 (2019).
31. Cadotte, D. W. et al. Characterizing the location of spinal and vertebral levels in the human cervical spinal cord. *AJNR Am. J. Neuroradiol.* **36**, 803–810 https://doi.org/10.3174/ajnr.A4192 (2015).
32. Bedard, S. & Cohen-Adad, J. Automatic measure and normalization of spinal cord cross-sectional area using the pontomedullary junction. *Front Neuroimaging* **1**, 1031253 https://doi.org/10.3389/fnimg.2022.1031253 (2022).
33. Sun, Z. et al. The genetic architecture of MRI derived human cervical spinal cord morphology reveals sensory-motor axis and biomarkers of neurological and systemic diseases. Preprint at *medRxiv* https://doi.org/10.1101/2025.08.05.25333008 (2025).
34. Dietrich, O., Raya, J. G., Reeder, S. B., Reiser, M. F. & Schoenberg, S. O. Measurement of signal-to-noise ratios in MR images: influence of multichannel coils, parallel imaging, and reconstruction filters. *J. Magn. Reson. Imaging* **26**, 375–385 https://doi.org/10.1002/jmri.20969 (2007).
35. Magnotta, V. A., Friedman, L. & First, B. Measurement of signal-to-noise and contrast-to-noise in the fBIRN Multicenter Imaging Study. *J. Digit. Imaging* **19**, 140–147 https://doi.org/10.1007/s10278-006-0264-x (2006).
36. Ganzetti, M., Wenderoth, N. & Mantini, D. Intensity inhomogeneity correction of structural MR images: a data-driven approach to define input algorithm parameters. *Front. Neuroinformatics* **10**, 10 https://doi.org/10.3389/fninf.2016.00010 (2016).
37. Atkinson, D., Hill, D. L. G., Stoyle, P. N. R., Summers, P. E. & Keevil, S. F. Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion. *IEEE Trans. Med. Imaging* **16**, 903–910 https://doi.org/10.1109/42.650886 (1997).
38. Hatamizadeh, A. et al. in *International MICCAI Brainlesion Workshop* 272–284 (Springer, 2021).
39. Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z. & Tu, Z. in *Artificial Intelligence and Statistics* 562–570 (PMLR).
40. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1607.08022 (2016).
41. Maas, A. L., Hannun, A. Y. & Ng, A. Y. in *Proc. ICML* 3.
42. Nesterov, Y. A method of solving a convex programming problem with convergence rate O (1/k** 2). *Dokl. Akad. Nauk SSSR* **269**, 543 (1983).
43. Sutskever, I., Martens, J., Dahl, G. & Hinton, G. in *International Conference on Machine Learning* 1139-1147 (PMLR).
44. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1711.05101 (2017).
45. Tan, M. X. & Le, Q. V. EfficientNet: rethinking model scaling for convolutional neural networks. In *Proc. Machine Learning Research* Vol. 97 (PMLR, 2019).
46. He, K., Zhang, X., Ren, S. & Sun, J. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778.
47. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1409.1556 (2014).
48. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
49. Somerville, L. H. et al. The Lifespan Human Connectome Project in Development: a large-scale study of brain connectivity development in 5-21 year olds. *Neuroimage* **183**, 456–468 https://doi.org/10.1016/j.neuroimage.2018.08.050 (2018).
50. Bookheimer, S. Y. et al. The Lifespan Human Connectome Project in Aging: an overview. *Neuroimage* **185**, 335–348 https://doi.org/10.1016/j.neuroimage.2018.10.009 (2019).
51. Sorenson, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analysis of vegetation on Danish commons. *Kong Dan. Vidensk. Selsk. Biol. Skr.* **5**, 1–5 (1948).
52. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 https://doi.org/10.2307/1932409 (1945).

53. Jaccard, P. The distribution of the flora in the alpine zone. 1. *New Phytol.* **11**, 37–50 (1912).
54. Huttenlocher, D. P., Klanderman, G. A. & Rucklidge, W. J. Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal.* **15**, 850–863 https://doi.org/10.1109/34.232073 (1993).

Acknowledgements

The authors gratefully thank all of the individuals for their participation in the study and the Friedreich's Ataxia Research Alliance (FARA, USA) for their help with FRDA participant recruitment. The authors thank and acknowledge FARA, Takeda Pharmaceuticals Company, Novartis Gene Therapies, IXICO plc, PTC Therapeutics, and Larimar Therapeutics for support in funding and for their participation in the TRACK-FA Neuroimaging Consortium, including providing advice and feedback on study design, implementation, and analysis of outcomes. Study data were collected and managed using REDCap electronic data capture tools hosted and managed by Helix (Monash University, Australia)[9,28]. Study data were collected, stored, and managed using XNAT, a software framework for managing neuroimaging laboratory data hosted by Monash (MXNAT) and Monash Biomedical Imaging (MBIXNAT) (Monash University, Australia). The authors thank the research coordinators across all TRACK-FA and READISCA sites for their assistance in participant recruitment and testing. The following collaborators opted to have their names acknowledged: Pedro Marcelino Mendes Novaes Melo, Department of Neurology, University of Campinas, Campinas, Brazil; Jennifer Farmer, Friedreich's Ataxia Research Alliance (FARA), Downingtown, Pennsylvania, United States of America; Tobias Linding, Department of Diagnostic and Interventional Neuroradiology, University Hospital Tübingen, Tübingen, Germany; Thomas H. Mareci, Department of Biochemistry and Molecular Biology, University of Florida, Gainesville, Florida, USA; Henry Paulsen, Department of Neurology, University of Michigan, Ann Arbor, MI, USA. Jörg B. Schulz, Department of Neurology, RWTH Aachen University, Aachen, Germany. JARA-BRAIN Institute Molecular Neuroscience and Neuroimaging, Research Center Jülich GmbH, Jülich, Germany; Caterina Tonon, Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy. IRCCS Institute of Neurological Sciences of Bologna, Functional and Molecular Neuroimaging Unit, Bologna, Italy.

Author contributions

Concept and design: G.M.J., D.H.S., I.H.H., T.J.R.R. Acquisition, analysis, or interpretation of data: G.M.J., D.H.S., I.M.A., T.A., O.G.P.B., B.B., F.C., L.A.C., M.C., L.C., E.D.A, A.D., M.B.D., I.D., A. Durr, G.F.E., J.F., J.F.R., J.V.G., W.G., N.G.K., E.V.H., P.G.H., C.R.H.-C., D.H., J.M.J., T.K., M.L., C.L., J.L., E.F.L., R.L., D.R.L., D.N.M., A.R.M.M., S.E.O., G.Ö., M.P., J.L.P., M.R., K.R., T.P.R., S.R., D.A.R, S.S., L.S., S.H.S., Z.S., M.S., H.A.G.T., G.D.V., B.V.d.W., Y.V.d.W., M.C.F., I.H.H., T.J.R.R. Drafting of the manuscript: G.M.J., D.H.S., I.H.H., T.J.R.R. Critical revision of the manuscript for important intellectual content: G.M.J., D.H.S., I.M.A., T.A., O.G.P.B., B.B., F.C., L.A.C., M.C., L.C., E.D.A., A.D., M.B.D., I.D., A. Durr, G.F.E., J.F., J.F.R., J.V.G., W.G., N.G.K., E.V.H., P.G.H., C.R.H.-C., D.H., J.M.J., T.K., M.L., C.L., J.L., E.F.L., R.L., D.R.L., D.N.M., A.R.M.M., S.E.O., G.Ö., M.P., J.L.P., M.R., K.R., T.P.R., S.R., D.A.R, S.S., L.S., S.H.S., Z.S., M.S., H.A.G.T., G.D.V., B.V.d.W., Y.V.d.W., M.C.F., I.H.H., T.J.R.R. Statistical analysis: G.M.J., I.H.H., T.J.R.R. Obtained funding: D.H.S., M.C.F., I.H.H., T.J.R.R. Supervision: I.H.H., T.J.R.R. All authors have read and approved the manuscript.

Funding

This work was supported by the Friedreich's Ataxia Research Alliance (FARA General Research Grant), São Paulo Research Foundation (FAPESP, grant #25/00186-4), the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES), and grants from the Australian National Health and Medical Research Council (NHMRC Ideas Grant 1184403). The funding agencies did not influence the study design, data collection, or manuscript drafting.

Competing interests

The authors declare no competing financial or non-financial interests: G.M.J., D.H.S., I.M.A., T.A., O.G.P.B., B.B., F.C., L.C., L.A.C., E.D.A, A.D., M.B.D., I.D., G.F.E., J.F.R., J.V.G., W.G., N.G.K., E.V.H., P.G.H., C.R.H-C., D.H., J.M.J., T.K., M.L., J.L., E.F.L., R.L., D.N.M., A.R.M.M., S.E.O., M.P., J.L.P., M.R., D.A.R., S.R., L.S., Z.S., M.S., H.A.G.T., G.D.V., B.V.d.W., Y.V.d.W., M.C.F. M.C. has received consulting fees from UniQure, Voyager, Entrada, TRiNDS and Neurogene and she is a co-founder of Ventura Life Sciences, LLC. A. Durr receives grants from the NIH, Foundation de la Recherche Médicale (FRM), Brain-Team and ANR. Paris Brain Institute receives consulting fees on A. Durr's behalf from Biogen, AskBio, Servier, UCB, Genome Quebec, and HuntIX. She holds partly Patent B06291873.5 «Anaplerotic therapy of Huntington's disease and other polyglutamine diseases. J.F. is a member of the of the European Reference Network for Rare Neurological Diseases and was funded within the Advanced Clinician Scientist Programme (ACCENT, funded by the Federal Ministry of Research, Technology and Space (BMFTR) under the funding code (FKZ): 01EO2107) and as a PI of the iBehave Network, sponsored by the Ministry of Culture and Science of the State of North Rhine-Westphalia and received funding from the National Ataxia Foundation (NAF); and has received consultancy honoraria from Vico therapeutics and Biogen, unrelated to the present manuscript. C.L. receives funding from the NIH, FDA, Department of Defense of the USA, FARA, NAF, and Biogen. D.R.L. Grants or contracts: NIH, FDA, Reata/Biogen, Larimar, PTC; royalties or licenses: patent on NMDA receptor encephalitis testing. G.Ö. received funding from the NIH, Biogen, National Ataxia Foundation (NAF) and has consulted for IXICO Technologies Limited, uniQure biopharma B.V., VICO Therapeutics, Servier, Sanofi and UCB Biopharma SRL/Lacerta Therapeutics Inc and serves on the Scientific Advisory Board of BrainSpec Inc. K.R. Grants or contracts from DFG, FARA, BMBF, Lilly, Eisai, Roche and Biogen. T.P.R. Consulting/equity arrangements with Prism Clinical Imaging, Proteus Neurodynamics, Fieldline Inc. and WestCan Proton Therapy Inc. S.S. FARA USA General Research Grant. S.H.S. Grants or contracts: NIH, FDA, MDA, Reata/Biogen, Avidity, Fulcrum, Amicus, Solid Biosciences, PTC, Arrowhead. Support for clinical trials: Reata, Biogen, PTC, Biohaven, Fulcrum, Avidity, Vertex, Arthex, Sanofi. Ad boards: Biogen, Fulcrum, Avidity, Solid, Amicus, PTC. I.H.H. Competitive or investigator-initiated research grants and contracts paid to employing institution, Friedreich Ataxia Research Alliance (USA), National Health and Medical Research Council (Australia), Telethon Fondazione (Italy), National Ataxia Foundation (USA), auDA Foundation (Australia), Steminent Biotherapeutics (Taiwan); consulting fees, Solid Biosciences (USA), Steminent Biotherapeutics (Taiwan), PTC Therapeutics (USA); travel support, National Ataxia Foundation (USA), Machado Joseph Disease Foundation (Australia). T.J.R.R. Research grant from São Paulo Research Foundation (FAPESP), Friedreich Ataxia Research Alliance, Biogen and NIH; travel support from Friedreich Ataxia Research Alliance, Biogen and NIH.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-03403-w.

Correspondence and requests for materials should be addressed to Thiago J. R. Rezende.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

Gustavo M. Jarola[1,45,46], Diogo H. Shiraishi[1,45], Isaac M. Adanyeguh[2], Tetsuo Ashizawa[3], Orlando Graziani Povoas Barsottini[4], Benjamin Bender[5], Fernando Cendes[1], Louise A. Corben[6,7,8], Manuela Corti[9], Léo Coutinho[10], Emile d'Angremont[11,12], Alfredo Damasceno[1], Martin B. Delatycki[7,8,13], Imis Dogan[14,15], Alexandra Durr[16], Gary F. Egan[17], Jennifer Faber[18,19,20], Juan Fernandez-Ruiz[21], Judith van Gaalen[22,23], William Gaetz[24], Nellie Georgiou-Karistianis[6], Eva van Heese[11,12], Pierre-Gilles Henry[2], Carlos R. Hernandez-Castillo[25], Diane Hutter[2], James M. Joers[2], Thomas Klockgether[18], Max Laansma[11,12], Christophe Lenglet[2], Jiakun Li[26], Eric F. Lock[26], Raffaele Lodi[27,28], David R. Lynch[29], David N. Manners[30], Alberto RM Martinez[1], Sergio Eiji Ono[31], Gülin Öz[2], Massimo Pandolfo[32], José Luiz Pedroso[4], Myriam Rai[33], Kathrin Reetz[14,15], Timothy P. Roberts[24], Sandro Romanzetti[14,15], David A. Rudko[32,34,35], Susmita Saha[6,36], Ludger Schoels[37,38], Sub H. Subramony[39], Zhuopin Sun[40], Matthis Synofzik[37,38], Hélio A. Ghizoni Teive[10,41], Gabriel de Deus Vieira[1], Bart van de Warrenburg[22], Ysbrand van der Werf[11,12], Marcondes C. França[1], Ian H. Harding[42,43,46], Thiago J. R. Rezende[1,44,46] ✉ & TRACK-FA Neuroimaging Consortium Isaac*

[1]Department of Neurology, School of Medical Sciences, University of Campinas (Unicamp), Campinas, Brazil. [2]Department of Radiology, Center for Magnetic Resonance Research, University of Minnesota, Minneapolis, MN, USA. [3]Stanley H. Appel Department of Neurology, Houston Methodist Research Institute and Weill Cornell Medicine, Houston, TX, USA. [4]Department of Neurology, Federal University of São Paulo, São Paulo, Brazil. [5]Department of Diagnostic and Interventional Neuroradiology, University Hospital Tübingen, Tübingen, Germany. [6]School of Psychological Sciences, The Turner Institute for Brain and Mental Health, Monash University, Melbourne, VIC, Australia. [7]Bruce Lefroy Centre, Murdoch Children's Research Institute, Parkville, VIC, Australia. [8]Department of Pediatrics, University of Melbourne, Parkville, VIC, Australia. [9]Department of Pediatrics, Powell Gene Therapy Centre, University of Florida, Gainesville, FL, USA. [10]Graduate program of Internal Medicine, Internal Medicine Department, Hospital de Clínicas, Federal University of Paraná, Curitiba, Brazil. [11]Department of Anatomy & Neurosciences, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [12]Neurodegeneration, Amsterdam Neuroscience, Amsterdam, The Netherlands. [13]Victorian Clinical Genetics Services, Parkville, VIC, Australia. [14]Department of Neurology, RWTH Aachen University, Aachen, Germany. [15]JARA-BRAIN Institute Molecular Neuroscience and Neuroimaging, Research Center Jülich GmbH, Jülich, Germany. [16]Sorbonne Université, Paris Brain Institute (ICM), Pitié-Salpêtrière Hospital, AP-HP, INSERM, CNRS, University Hospital Pitié-Salpêtrière, Paris, France. [17]Monash Biomedical Imaging, Monash University, Melbourne, VIC, Australia. [18]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany. [19]Department of Parkinson, Sleep and Movement Disorders, Center of Neurology, University Hospital Bonn, Bonn, Germany. [20]Department of Neuroradiology, University Hospital Bonn, Bonn, Germany. [21]Neuropsychology Laboratory, Department of Physiology, Faculty of Medicine, National Autonomous University of Mexico, City of Mexico, Mexico. [22]Department of Neurology, Donders Institute for Brain, Cognition, and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands. [23]Department of Neurology, Rijnstate Hospital, Arnhem, The Netherlands. [24]Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [25]Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada. [26]Division of Biostatistics and Health Data Science, School of Public Health, University of Minnesota, Minneapolis, MN, USA. [27]Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy. [28]Functional and Molecular Neuroimaging Unit, IRCCS Institute of Neurological Sciences of Bologna, Bologna, Italy. [29]Department of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [30]Dipartimento di Scienze per la Qualità della Vita, University of Bologna, Bologna, Italy. [31]Clínica DAPI – Diagnóstico Avançado Por Imagem, Curitiba, Brazil. [32]Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada. [33]Friedreich's Ataxia Research Alliance, Downigtown, PA, USA. [34]McConnell Brain Imaging Centre, Montreal Neurological Institute and Hospital, Montreal, QC, Canada. [35]Department of Biomedical Engineering, McGill University, Montreal, QC, Canada. [36]Department of Neuroscience, School of Translational Medicine, Monash University, Melbourne, VIC, Australia. [37]Department of Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, Tübingen, Germany. [38]German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany. [39]Department of Neurology and the Fixel Institute for Neurological Diseases, University of Florida, Gainesville, FL, USA. [40]Genetics and Gene Regulation Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia. [41]Movement Disorders Unit, Neurology Service, Internal Medicine Department, Hospital de Clínicas, Federal University of Paraná, Curitiba, Brazil. [42]QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia. [43]School of Translational Medicine, Monash University, Melbourne, VIC, Australia. [44]Department of Neurosciences, School of Medicine, University of São Paulo at Ribeirão Preto (USP-RP), Ribeirão Preto, Brazil. [45]These authors contributed equally: Gustavo M. Jarola, Diogo H. Shiraishi. [46]These authors jointly supervised this work: Ian H. Harding, Thiago J. R. Rezende. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: thiago.jrezende@gmail.com

TRACK-FA Neuroimaging Consortium Isaac

Isaac M. Adanyeguh[2], Louise A. Corben[6,7,8], Manuela Corti[9], Martin B. Delatycki[7,8,13], Imis Dogan[14,15], William Gaetz[24], Nellie Georgiou-Karistianis[6], Pierre-Gilles Henry[2], James M. Joers[2], Christophe Lenglet[2], Jiakun Li[26], Eric F. Lock[26], David R. Lynch[29], Alberto RM Martinez[1], Massimo Pandolfo[32], Myriam Rai[33], Kathrin Reetz[14,15], Timothy P. Roberts[24], Sandro Romanzetti[14,15], David A. Rudko[32,34,35], Susmita Saha[6,36], Sub H. Subramony[39], Marcondes C. França[1], Ian H. Harding[42,43,46] & Thiago J. R. Rezende[1,44,46] ✉

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.