Effects of variability in manually contoured spinal cord masks on fMRI co-registration and interpretation

Mark A. Hoggarth[1]*, Max C. Wang[2]*, Kimberly J. Hemmerling[1,2], Andrew D. Vigotsky[2,3], Zachary A. Smith[4], Todd B. Parrish[2,5], Kenneth A. Weber II[6], Molly G. Bright[1,2]

Affiliations:

1. Department of Physical Therapy and Human Movement Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
2. Department of Biomedical Engineering, McCormick School of Engineering, Northwestern University, Evanston, IL, USA
3. Department of Statistics, Weinberg College of Arts and Sciences, Northwestern University, Evanston, IL, USA
4. Department of Neurological Surgery, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA
5. Department of Radiology, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
6. Systems Neuroscience and Pain Lab, Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University, Palo Alto, CA, USA

*Joint First Authors

Corresponding Author:
Mark A. Hoggarth PT, DPT, PhD
645 N. Michigan Ave.
Suite 1100
Chicago, IL 60611
312-503-5870
Mark.hoggarth@northwestern.edu

Abstract

Functional magnetic resonance imaging (fMRI) of the human spinal cord (SC) is an emerging but technically challenging method for understanding neurovascular responses to sensory and motor stimuli. Group-analysis of SC fMRI data involves co-registration of subject-level data to a standard image, which currently requires manual masking of the cord. We hypothesize that this manual input may contribute to bias or poor repeatability in group-level SC fMRI results. To test this, we examined variability in SC masks drawn in fMRI data from 21 healthy participants taken from a completed study mapping responses to sensory stimuli of the C7 dermatome. Masks were drawn on the temporal mean functional image by eight raters with varying levels of neuroimaging experience, and the rater from the original study acted as an expert reference. Spatial agreement between rater and expert masks was measured using the Dice Similarity Coefficient (DSC), and the influence of rater experience and individual dataset was examined using ANOVA. Each rater's masks were then used for spatial normalization of functional imaging data to the PAM50 template image. Using tissue regions defined in the template image, the gray matter-white matter signal contrast of the aligned functional data was used to evaluate the relative accuracy of the spatial normalization procedure across raters. Subject- and group-level analysis of activation during the left- and right-sided sensory stimuli were performed for each rater's co-registered data. Agreement with the expert's SC mask was associated with both rater experience level ($F(7,140) = 32.12$, $P < 2\times10\text{-}16$, $\eta^2 = 0.29$) and dataset ($F(20,140) = 20.58$, $P < 2\times10\text{-}16$, $\eta^2 = 0.53$). Dataset variations may reflect image quality metrics: the ratio between the signal intensity of spinal cord voxels and surrounding cerebrospinal fluid was correlated with DSC results ($p<0.001$). As predicted, variability in the manually-drawn masks visibly influenced spatial normalization, and the GM:WM contrast in the registered data showed significant effects of rater and dataset (rater: $F(8,160) = 23.57$, $P < 2\times10^{-16}$, $\eta^2 = 0.24$; dataset: $F(20,160) = 22.00$, $P < 2\times10^{-16}$, $\eta^2 = 0.56$). These registration differences were propagated into activation mapping results, and subject-level activation maps showed rater-dependent agreement with the expert's maps. Although group-level activation maps differed between raters, no systematic bias was identified. While increasing consistency in manual contouring of spinal cord fMRI data could improve data co-registration and ultimately the inter-rater agreement in activation mapping, our results suggest that other improvements in image acquisition and post-processing may be more critical to address.

Keywords:

Functional MRI, Spinal cord, Contouring, Image Registration, Reliability

Introduction

Functional magnetic resonance imaging (fMRI) of the spinal cord is an emerging technique for understanding neurovascular responses to sensory and motor stimuli.[1] However, analysis and interpretation of spinal cord fMRI data is hindered by the small size of the cord and physiological noise from cardiac and respiratory motion.[2,3] Combining fMRI data from many individuals for group-level analysis is one important way to improve sensitivity to neural activation and further understand the neuroanatomical structure of the spinal cord. Central to this challenge, current image co-registration techniques in spinal cord fMRI recommend manual masking of the cord when registering functional images to a standard space.[4] Image noise, insufficient spatial resolution, and poor contrast between spinal cord tissue and the surrounding cerebrospinal fluid (CSF) may lead to subjective differences when contouring the cord, which could result in systematic bias in data co-registration and thus alter individual and group analysis results.

Differences in functional image processing pipelines of the brain have been shown to have large potential impacts on the resultant findings of a study.[5] Additionally, while there are multiple software packages for spatial normalization of fMRI in the brain[6-8], spinal cord studies have fewer options or must create bespoke techniques for registration to labeled structural templates of the cord.[4,9,10] Bowring et al. observed that variability in results from task fMRI in the brain are heterogeneous, depend on the input dataset, may result from each aspect of an image processing pipeline (including registration), and also are impacted by the software package utilized.[5] Given the lack of a unified method of preprocessing spinal cord fMRI data, reproducibility of a reported results is likely to be poor, and improvements are needed to standardize image processing and reduce sources of variability and bias at every stage of analysis.

In this work, we examine the effects of mask variability at different stages of a single analysis pipeline for spinal cord fMRI using a previously published study dataset. We characterize the variability in spinal cord masks achieved by eight raters with varying levels of image analysis experience, with respect to an “expert” rater from the original study and publication. We then demonstrate how this variability is propagated following registration of functional imaging data to a standard spinal cord template image. We subsequently run individual- and group-level analyses for a sensory stimulus task, using each rater’s masks during co-registration, to assess the impact on fMRI activation patterns at the single-subject and overall study level. Finally, we discuss the causes of this variability, their importance, and make recommendations for prioritizing future improvements in spinal cord fMRI.

Methods

2.1 Image acquisition and experimental protocol

This work utilized a subset of anatomical and functional MRI data from 21 healthy participants from a previous study.[11] Images were acquired using a 3T Siemens Prisma scanner (Siemens, Erlangen, Germany), utilizing a 64-channel head/neck coil and a SatPad™ cervical collar (SatPad Clinical Imaging Solutions, West Chester, PA, USA). Anatomical T2-weighted images were acquired covering the cervical and upper thoracic spine, using the SPACE sequence (Siemens, Erlangen, Germany), with parameters: TR = 1500ms, TEeff = 135 ms, echo train = 74m flip angle = 90°/140°, slices = 64, effective voxel size = 0.8 x 0.8 x 0.8 mm^3, iPAT acceleration factor = 3, interpolated in-plane resolution = 0.4 x 0.4 mm^2.[12,13] T2*-weighted functional scans of the cervical spinal cord were acquired, centered at the C5 vertebral level, using a gradient-echo echo-planar-imaging sequence with ZOOMit selective field-of-view.[14-16] Functional imaging parameters were: TR_{3D} = 2000 ms, TE = 30 ms, flip angle = 80°, volumes = 450, slice order was interleaved, field-of-view = 128 × 44 mm^2, acquisition matrix = 128 × 44 voxels, in-plane resolution = 1 × 1 mm^2, slice thickness = 3 mm, two dummy volumes discarded.

During each functional scan, alternating left and right tactile stimuli were applied to the dorsum of each hand in the C7 dermatomal region.[11] Stimulation was applied manually at approximately 2 Hz by examiners in the scan room. Stimulation lasted for 15 sec on each hand and was interspersed with a 15 second rest period.

Preprocessing of the functional MRI time series was performed as in the original study. Motion correction was performed using FMRIB's Linear Registration Tool (FLIRT) for a rigid and non-rigid 3D alignment, followed by rigid 2D (axial slice) alignment.[7,17] Temporal mean images were calculated from these preprocessed functional data. FMRI data were further denoised by removing periodic physiological noise confounds (RETROICOR)[18,19], warped to the PAM 50 spinal cord template, and then smoothed with a $2mm^3$ full width half maximum Gaussian smoothing kernel prior to group analysis.[20]

2.2 Manual contouring

Raters (N=8) with differing levels of functional neuroimaging experience were recruited to manually contour the spinal cord on the temporal mean functional images in the axial plane from all 21 study participants. Raters were categorized by their functional imaging experience: Proficient (N=1, >10 years), Competent (N=4, 1–9 years), and Novice (N=3, <1 year). Note that experience was not specifically focused on spinal cord functional imaging, and therefore our rater categories reflect overall time in the imaging field rather than deep expertise in spinal cord anatomy and imaging artifacts. This limitation is discussed in a later section. Additionally, as this

dataset was collected for a previously completed and published study,[11] the original masks from that work were included as "Expert" reference contours.

2.3 Registration to standard space

Image registration was performed with the Spinal Cord Toolbox.[9] First, the sct_deepseg_SC function was implemented to identify the cord in the high-resolution T2-weighted anatomical images[21], and the C3 and C7 vertebrae were manually labeled. The spinal cord segmentation and vertebral level labels were then used to register the anatomical image to the PAM50 spinal cord template image using the sct_register_to_template function.[22] The anatomical-to-template registration was performed once for each dataset and did not vary between raters. The temporal mean functional images were registered to the PAM50 template using the sct_register_multimodal function, utilizing each rater's manually contoured masks of the cord and initial warping field (generated from the anatomical registration above) as inputs. An overview of the masking and registration process is shown in **Figure 1**.

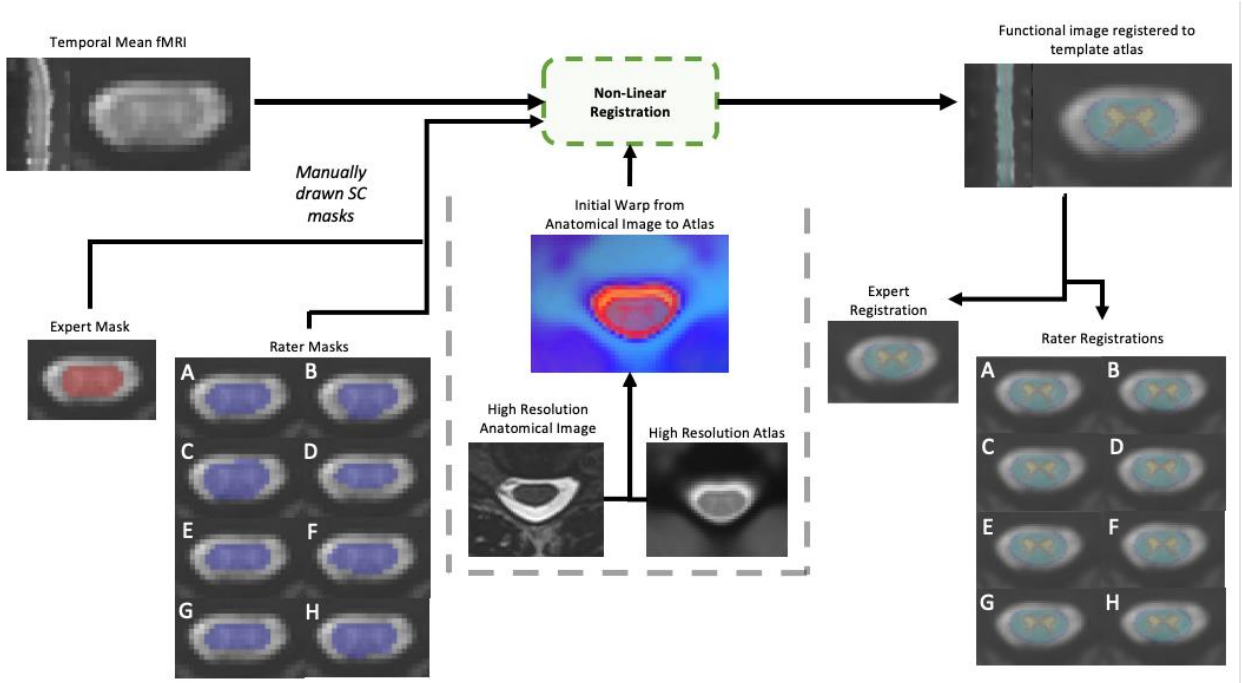


Figure 1: *Outline of non-linear registration. A mean temporal image was created; then an expert and 8 raters with varied levels of experience contoured the spinal cord manually. Non-linear registration was performed with the Spinal Cord Toolbox, utilizing additional information from a high resolution anatomical T2 weighted image. The impact of the resulting registration of the functional image to template space was analyzed.*

2.4 Variability in pre- and post-registration masks

The variability in rater-contoured masks was assessed by comparing each rater mask (RM) to the expert mask (EM). To quantify differences, the Dice Similarity Coefficient (DSC) was calculated for the total volume and for each axial slice as:

$$DSC = \frac{2*|EM \cap RM|}{|EM|+|RM|}.$$

ANOVA was used to compare DSC, averaged by dataset, between raters and the expert, with a significance level α=0.01. The most superior and inferior slices were not included in these analyses due to poor image quality.

To assess the effect of fMRI image quality on rater agreement, two metrics were calculated for each dataset on the temporal mean functional image: 1) the coefficient of variation (CV) of voxels adjacent to the expert spinal cord mask and 2) the ratio between the mean signal of adjacent voxels and the mean signal within the expert's spinal cord mask (Adjacent:SC ratio). For both metrics, adjacent voxels were defined by a 2-voxel dilation from the expert SC mask. The CV of adjacent voxels captures inconsistencies in CSF signal and the Adjacent:SC ratio represents the signal contrast required to contour spinal cord boundaries. A higher CV or lower Adjacent:SC ratio could obscure the delineation between spinal cord and surrounding CSF and would increase the difficulty of contouring the cord, potentially leading to increased variability in rater masks.

DSC and fMRI image quality metrics were then compared for each imaging slice across all participant datasets. When raters contour spinal cord images, the cord is primarily viewed in the axial plane. In this orientation, image quality can vary across slice acquisitions: for example, there is decreased efficiency of head/neck coils in more inferior slices, where magnetic field inhomogeneities may also be increased due to magnetic susceptibility variation during respiration.[23] While shimming of the magnetic field can reduce static field inhomogeneities, present methods are still incapable of fully compensating for smaller field variations due to changes in anatomical structures such as the borders between spinal discs and vertebral bodies[24], and custom dynamic shimming techniques are not routinely available to fully mitigate dynamic effects of respiration Additionally, under normal breathing conditions, the spinal cord moves most in the C4-T1 region.[25] Combined, these factors can create significant variation in the image properties across axial slices of a given data set. We therefore calculated correlation coefficients between both slicewise metrics of image quality and log-transformed DSC values, for each RM (with respect to the EM) for all datasets.

To evaluate the accuracy of the alignment of functional data to the standard PAM50 template space, we considered the signal contrast between gray matter (GM) and white matter (WM) voxels

using regions defined by the PAM50 spinal cord atlas. Inherent signal contrast between GM and WM tissues within the spinal cord will vary across acquisitions, and the functional BOLD-weighted sequence was not designed to optimize GM:WM contrast; therefore, this value is likely small and highly variable across datasets. However, we use this metric to evaluate the relative accuracy of the rater-specific spatial normalization procedures for a given dataset: perfect alignment with the PAM50 template should lead to maximal GM:WM contrast using regions defined by that template atlas. Imperfect alignment with the PAM50 template would cause mixing of GM and WM signals across the two atlas regions, reducing the observed GM:WM contrast for that dataset. The GM:WM ratio was calculated for each dataset, following spatial normalization procedures using each rater's spinal cord mask (or the expert's). These values were log-transformed, then fitted via an ANOVA with rater and dataset as independent variables, thus allowing us to distinguish the impact of rater on spatial normalization accuracy from inherent variability in signal contrast across the datasets.

2.5 Variability in statistical activation in participant- and group-level analyses

As described in the original study, participant-level analyses were performed to characterize significant activation associated with the sensory stimuli.[11] Trialwise left- and right-sided stimuli were convolved with hemodynamic response function and analyses were performed via a generalized linear model, using FILM with prewhitening.[26] Voxels with a Z-score > 2.3 ($p < 0.01$, uncorrected) were classified as active. These analyses were repeated for each rater. Participant-level activation maps (z-statistics) for left- and right-sided sensory stimuli were calculated using a fixed-effects analysis for each dataset and rater. Based on spinal cord anatomy, activation from the tactile stimulus was expected to localize to the ipsilateral hemicord in the area of the C7 spinal level. We therefore initially focused our attention on ipsilateral activations, only. The spinal cord was divided into the left and right hemicord, excluding from analysis the center column of spinal cord voxels where the hemicords meet. We considered activation with the left-sided stimulus in the left hemicord, and activation with the right-sided stimulus in the right hemicord, calculating the spatial correlation (Fisher's z) between the ipsilateral activation patterns for each rater and the expert rater for each individual dataset.

Group-level activation results were achieved using the participant-level activation maps derived from each rater (and the expert). As in the original study, all group-level analyses were performed in the region of intersection of the functional images, again using a fixed-level analysis in FILM, where voxels with a Z-score > 2.3 and multiple comparisons correction cluster significance threshold of $p < 0.05$.[26] The original study found that activity was somewhat lateralized to the ipsilateral cord, but also deviated from these expectations, identifying activity more broadly distributed throughout the dorsal and ventral aspects of the cord and also superiorly and inferiorly to the expected cervical level.[11] Thus, considering the left and right hemicord regions separately, we assessed the distribution of z-statistics associated with ipsilateral activation (left cord, left

stimulation; right cord, right stimulation) and contralateral activation (left cord, right stimulation; right cord, left stimulation) in the group-level results. These distributions were compared between the expert and the raters by standardizing (i.e., *z*-scoring) each of the expert's *z*-statistics (y_i) relative to the raters' distribution ($\overrightarrow{x_i}$) for each voxel $i = 1, \ldots, n$, and then averaged the results:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i - \overline{x_i}}{\sigma_{x_i}},$$

Similarly, the standard deviation was also calculated:

$$\mathrm{SD}[z] = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i - \overline{x_i}}{\sigma_{x_i}} - \bar{z}\right)^2}.$$

Results

DSC comparisons between individual raters and the expert demonstrated trends with both dataset and rater experience, shown in **Figure 2,** where datasets are sorted by average DSC (rater: $F(7,140) = 32.12$, $P < 2\times10\text{-}16$, $\eta^2 = 0.29$; dataset: $F(20,140) = 20.58$, $P < 2\times10\text{-}16$, $\eta^2 = 0.53$). While the vertical trend in **Figure 2** shows that DSC was typically higher for the Proficient-level rater and decreased for Competent and Novice levels (i.e., the Proficient rater created spinal cord masks most similar to the Expert Rater), the horizontal trend illustrates differences in DSC attributable to dataset features, potentially including image quality or subject anatomy. Reported in **Table 1**, the CV of voxels adjacent to the spinal cord was not strongly correlated with DSC for most raters (6 of 8), and demonstrated both positive and negative correlations, with coefficients ranging from -0.114 to 0.139. However, the Adjacent:SC signal ratio was correlated with rater DSC in all cases ($p<0.001$), with coefficients ranging from 0.227 to 0.566. All corresponding Adjacent:SC and DSC values, for each imaging slice of each dataset, are shown in **Supplementary Figure 1**.

Qualitatively, the registration of functional images to template space and the inverse registration of the template atlas to the functional image both showed visible differences in alignment (**Figure 3**). For example, in Dataset 2, shown with Expert and Novice registrations, the PAM50 template masks of GM and WM are clearly not co-localized.

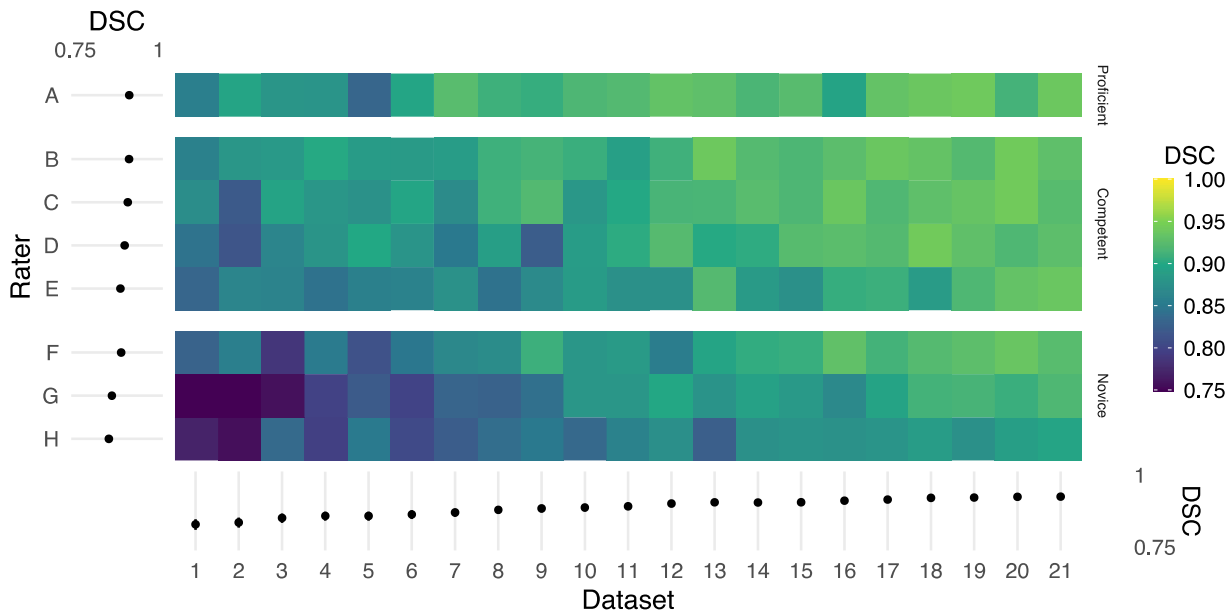


Figure 2: *Color map of Dice similarity coefficients (DSC) comparing each rater mask with the expert mask for each of 21 datasets. Raters are organized from lowest to highest average DSC and segmented into 3 categories based on their functional imaging experience: A: Proficient (N=1, >10 years), B–E: Competent (N=4, 1–9 years), and F–H: Novice (N=3, <1 year). Datasets are organized from lowest to highest average DSC. The margins (left and bottom) contain estimated marginal means for each rater and dataset, along with their 95% confidence intervals. Horizontal and vertical gradient trends indicate the effect of rater experience and dataset, respectively, on agreement with the expert.*

Table 1. *Pearson's Correlations Between Rater agreement (DSC) with Expert and Coefficient of Variation (CV) and Adjacent:SC Ratio.*

	DSC vs. CV			**DSC vs. Adjacent:SC**		
Rater	**Pearson's *r* ($CI_{95\%}$)**		***p*-value**	**Pearson's *r* ($CI_{95\%}$)**		***p*-value**
A	0.10	(0.01, 0.18)	0.029	0.23	(0.14, 0.31)	**< .001**
B	0.12	(0.04, 0.21)	**0.005**	0.47	(0.39, 0.53)	**< .001**
C	0.14	(0.05, 0.22)	**0.001**	0.35	(0.27, 0.43)	**< .001**
D	−0.10	(−0.19, −0.01)	0.028	0.26	(0.17, 0.34)	**< .001**
E	−0.04	(−0.13, 0.04)	0.325	0.42	(0.34, 0.49)	**< .001**
F	0.08	(−0.01, 0.16)	0.086	0.43	(0.36, 0.50)	**< .001**
G	−0.11	(−0.19, −0.02)	0.014	0.57	(0.50, 0.62)	**< .001**
H	−0.11	(−0.20, −0.03)	0.01	0.29	(0.21, 0.37)	**< .001**

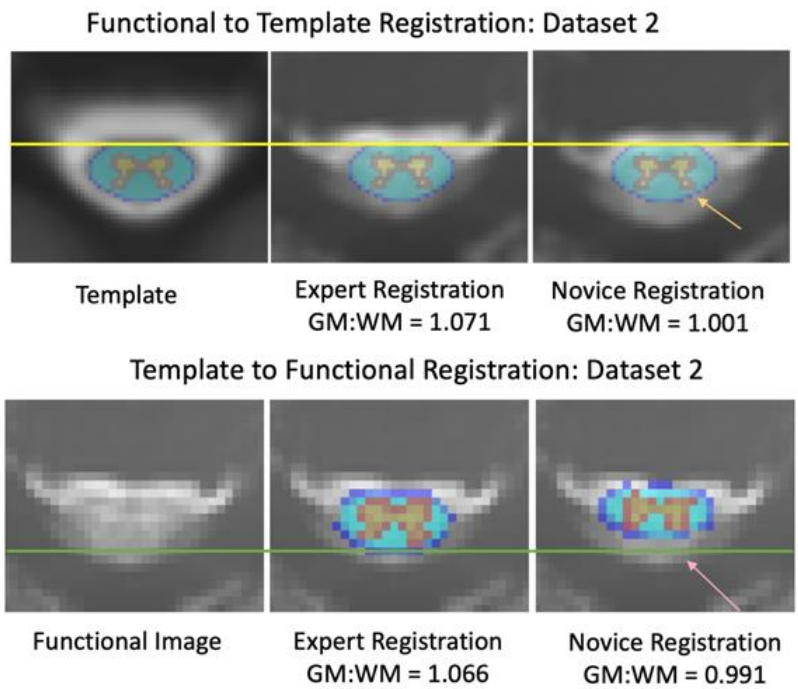


Figure 3. *Differences in alignment between an expert and a novice registration, from dataset 2. The mean DSC was 0.736 across the axial slices for the input masks between the expert and rater for dataset 2. (**Top**) Registration of functional images to the PAM50 template. The yellow line represents the most anterior white matter voxel coordinate of the template mask. The orange arrow indicates an area of grey matter in the dorsal horn that is not aligned with the template mask in a registration with novice masking. (**Bottom**) Registration of the PAM50 template atlas to functional image space. The green line represents an estimate of the most dorsal coordinate of the functional image. The pink arrow indicates an area of grey matter in the dorsal horn that is not aligned with the atlas in novice registration. Accuracy of registration alignment is supported by the GM:WM ratio, as misalignment introduces a mixing of GM, WM, and potentially CSF voxels into the masked areas, reducing the observed contrast.*

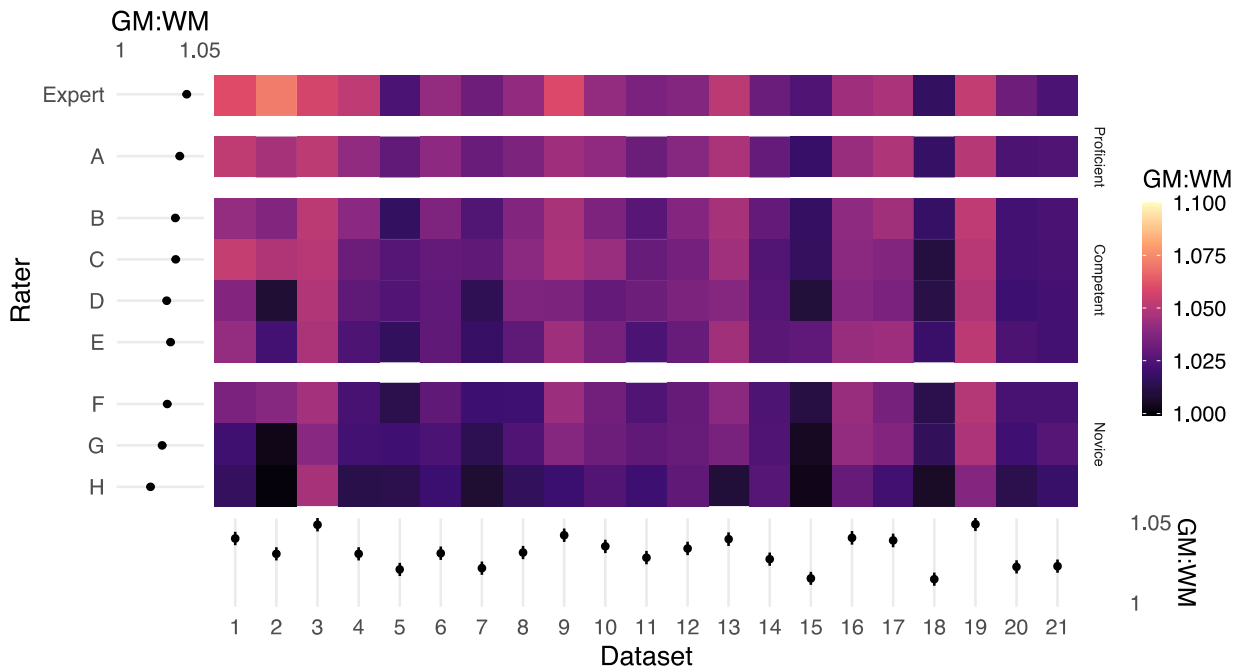


Figure 4. *Color map of GM:WM contrast across all rater registrations. Higher GM:WM contrast for a given input dataset indicates relatively better registration alignment. The margins (left and bottom) contain estimated marginal means for each rater and dataset, along with their 95% confidence intervals. Rater experience is correlated with higher levels of GM:WM contrast following spatial registration. Moreover, there are marked dataset effects on GM:WM, which are reflected by the ANOVA results.*

GM:WM contrast, calculated using the PAM50 template masks following spatial normalization of the functional data to the PAM50 template image (akin to the top panels in **Figure 3**), increased with rater experience within datasets, illustrated by a vertical gradient (**Figure 4**). As anticipated, there is also substantial variability of GM:WM contrast across the individual datasets, unrelated to rater masking and normalization. ANOVA revealed that both rater and dataset had marked contributions to GM:WM variance, with dataset contributing relatively more variance (rater: $F(8,160) = 23.57$, $P < 2\times10^{-16}$, $\eta^2 = 0.24$; dataset: $F(20,160) = 22.00$, $P < 2\times10^{-16}$, $\eta^2 = 0.56$) as anticipated.

Fisher's *z* spatial correlations between individual participant ipsilateral activation maps generated by each rater and the expert are shown in **Figure 5**. Horizontal and vertical trends are both present in the spatial correlations, indicating dataset and rater effects are factors in observed activation maps at the individual-level. Note, datasets were ordered by initial DSC agreement (as described for **Figure 2**) for this visualization. Results from group-level analyses are also shown in **Figure 5** and **Table 2,** where results from z-testing between the expert and raters for each stimulus and

hemicord condition (*e.g.*, Left hemicord, left stimulus) were distributed about 0 for all cases, illustrating no average Z-score differences. Detailed in **Table 3**, spatial correlations between the raters' and expert's group-level results ranged from 0.954 to 0.875 (mean of 0.923) for the ipsilateral activations, and 0.952 to 0.781 (mean of 0.892) for the contralateral activations. Thus, while systematic trends were not observed across the results of different raters, there was observable disagreement in activation mapping at the group-level. An illustration of these differences is shown in **Supplemental Figure 2**.

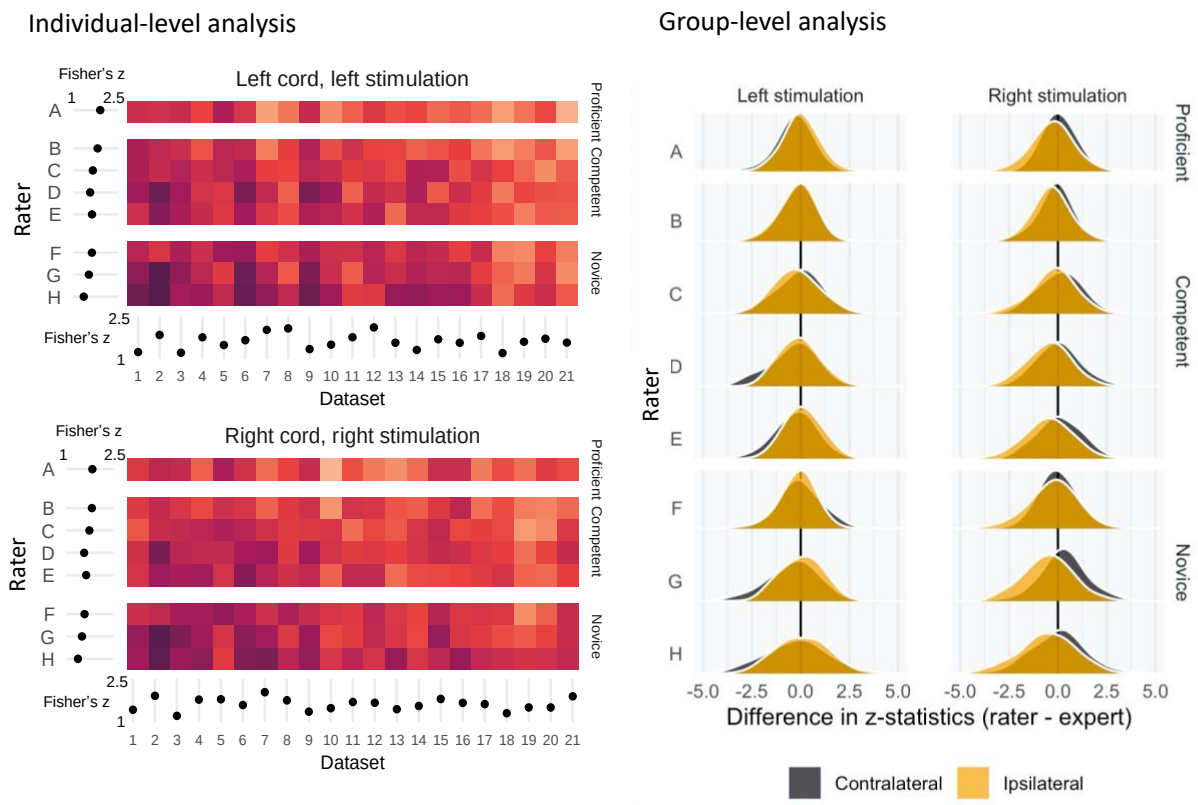


Figure 5: *Summary of the individual- and group-level activation maps. (**Left**) Spatial correlation of statistical activation maps derived from the individual-level analysis results of each rater and the expert, for ipsilateral activation of left- and right-sided stimulation trials. Results shown as a color map of Fisher's z correlations between rater and expert statistical activation maps, with the dataset order matching Figures 2 and 4. The horizontal and vertical gradient trends indicate that dataset and experience influence agreement with the expert's individual-level statistical activation maps. Marginal mean Fisher's z scores by data set and rater are shown at the axes. (**Right**) Distribution of group-level activation z-statistics in the left and right hemicords for left- and right-sided stimulation trials. The shape and centering of the distributions are generally similar across all raters illustrating no systematic difference in average z-score, however, there was observable spatial disagreement throughout the activation maps when comparing the expert to rater's group results. Example group-level activation maps aer shown in Supplemental Figure 2.*

***Table 2**: Magnitude of Expert's group-level z-statistics relative to the raters' distribution. The expert's group-level z-statistic for each voxel was scaled by the distribution of z-statistics from the other raters. Presented here are the mean and standard deviations of the standardized Expert z-statistics. A mean of 0 would indicate that the Expert has the same z-statistic as the other raters (on average); values greater than zero indicate greater z-statistics than the raters' average; and values lower than zero indicate lower z-statistics than the raters' average.*

Side	Stimulation	Mean	± SD
Left	Ipsilateral	0.05	±1.07
Right	Ipsilateral	0.30	±1.35
Left	Contralateral	-0.09	±1.27
Right	Contralateral	0.49	±1.30

Table 3: *Group-level activation map spatial correlations (Fisher's z) between each rater and the expert by side(Left (L) or Right (R)) and stimuli condition (ipsilateral or contralateral activation).*

		Spatial Correlation (Fisher's *z*)	
Rater	**Side**	**Ipsilateral**	**Contralateral**
A	*L*	0.950	0.952
	R	0.931	0.924
B	*L*	0.952	0.942
	R	0.954	0.935
C	*L*	0.916	0.902
	R	0.930	0.871
D	*L*	0.933	0.901
	R	0.923	0.863
E	*L*	0.943	0.907
	R	0.905	0.896
F	*L*	0.944	0.930
	R	0.910	0.895
G	*L*	0.928	0.890
	R	0.902	0.839
H	*L*	0.875	0.849
	R	0.877	0.781

Discussion

Mask differences were attributable to both rater experience and dataset quality. Because these differences naturally occur along the edges of the cord, at the anatomical boundary of spinal cord white matter and surrounding CSF, we hypothesized that the signal contrast between these regions in each dataset would be associated with the agreement of rater masks with the expert. Indeed, better contrast at the edges of the cord (higher Adjacent:SC ratio) was significantly correlated with higher values of DSC, shown in **Table 1**. **Supplementary Figure 1** illustrates the slice-wise DSC and Adj:SC ratio correlations for every rater. Interestingly, we observe many datasets in this study presented with Adjacent:SC ratios less than 1, which is not anticipated given the T2* of tissue versus CSF. Furthermore, this ratio shows added variability across longitudinal image slices of each dataset. The breakdown of this expected contrast was often observed in the dorsal aspect of the cord, perhaps due to a posterior shift while a participant is being scanned in a supine position.[27] A posterior shift of the cord reduces the amount of CSF buffer between neural tissue and ligamentum flavum, disc, or bone from the spinal canal and may lead to increased partial volume averaging of these tissues. CSF flow may also impact the signal intensity of CSF voxels in this acquisition, potentially increasing or decreasing voxel brightness, thus influencing the Adjacent:SC ratio and mask fidelity. This study used the temporal mean functional image for contouring the spinal cord, which may merge these flow artifacts and reduce apparent image contrast. It may be more successful to contour the spinal cord on one fMRI volume, rather than the temporal mean image, if a volume with maximal Adjacent:SC contrast can be identified. However, it will be challenging to do this in a robust and systematic manner and appropriately integrate this step into volume realignment (i.e., motion correction) procedures, and this would not fully compensate for inherently poor Adjacent:SC contrast across the scan. These findings support the need for continued improvement in spinal cord fMRI acquisition techniques to improve and stabilize image contrast along the length of the cord while mitigating flow artifacts, such as improved receive coils[4,28-9], higher static magnetic field strengths[4,29], sequence and protocol optimization[4,15-6], and image processing techniques [2-4,9,28].

Following spatial normalization to a template image, the functional image GM:WM ratio (using GM and WM masks from the PAM50 template) was also reflective of rater expertise and image quality. In T2*-weighted images, GM is expected to be brighter than WM, and thus a well-registered functional dataset will yield a more positive GM:WM contrast compared to a less successfully registered version of the same input data. In our results (**Figure 4**), disagreement with the expert in masking appeared associated with lower GM:WM ratios following image registration, as the underlying tissue projections onto the template contained a mix of tissue classes. The expert achieved the highest GM:WM values, greater than any of the other raters. These results show that rater experience in neuroimaging influences the success of image registration to template space, supporting the value of training for any individual manually contouring the spinal cord in fMRI

studies. However, reducing the dependency of the image registration algorithms on manual inputs could potentially remove this confound: as has been done for the registration of high-resolution anatomical images of the spinal cord, convolutional neural networks could be trained for automated cord segmentation in fMRI datasets.[21] Such advances would have the added benefit of speeding up the image processing pipeline by removing the rate-limiting step of manual contouring.

Differences in masking were shown to impact co-registration of individual datasets and the spatial distribution of activation at the individual-level, shown in **Figure 5 (left)**, however there was not a clear systematic impact on average group-level activation maps, as shown in **Figure 5 (right)** This finding may be due to averaging over a relatively large voxel size (1x1x3mm) when evaluating the cord with an average size of 7.4±0.9mm anterior-posterior and 11.4±1.2mm left-right at the C7 level.[30] Additionally, the majority of BOLD activation is expected near the center of the cord, in GM, where rater masks are most likely to agree, and the nonlinear registration algorithm may produce more consistent results. **Supplementary Figure 3** illustrates the location of differences in masking between the expert and rater masks.

However, there did appear to be a positive effect of rater experience on the spatial correlation between their group-level activation map and the expert's. Also, as shown in **Supplementary Figure 2**, although there were no average differences in mean Z-scores, the regions of activity varied between all raters. Furthermore, because we observe varying GM:WM contrasts in the co-registered data, this suggests that variations at the edges of rater masks can influence voxel positioning in more central regions of the cord following spatial normalization.

Still, the lack of systematic differences in the group-level results may also be due to the inherent challenges of measuring BOLD responses in the spinal cord, where there are poorly resolved physiologic motion confounds and a small anatomical target relative to the image resolution. In the original work with expert masking, the classical definitions of dermatomal sensory distributions were not clearly observed in the group-level activation to left and right sensory stimuli.[11] Although predominantly ipsilateral activations were observed, they were not localized to the dorsal aspects of the cord, and activation spread across vertebral levels rather than localizing to the expected C7 region. One interpretation is that there is a more complex anatomical network underlying spinal neurological response to stimuli, and we refer the reader to the original work for a more detailed neurophysiological interpretation. However, it could also suggest that there remains a fundamental obstacle of low signal-to-noise ratio in spinal cord fMRI that hinders robust mapping of true activations. The results of this study indicate that such limitations in spinal cord fMRI sensitivity and specificity may be more critical to activation mapping than the subtle variations in image co-registration that arise from manually contoured spinal cord masks. However, as image quality improves, through developments in hardware, acquisition strategies,

and image processing techniques, it may become apparent that co-registration of functional data to a standard template space is increasingly important in achieving accurate and robust group-level results.

Conclusion

We observed differences in individual rater masks of the spinal cord in fMRI data when compared to masks from an expert rater. These differences were driven by both the neuroimaging experience level of the rater and the individual dataset image quality and led to variable co-registration with a standard spinal cord template image. This variability propagated into differences in individual-level fMRI activation results, as measured via spatial correlation between the expert and raters' activation maps for left and right sensory stimuli. However, when performing group-level analyses, these masking and co-registration differences did not have a systematic effect on the average Z-score of resulting group-level activation. While increasing consistency in manual contouring of spinal cord fMRI data could improve data co-registration and ultimately the inter-rater agreement in activation mapping, our results suggest that other improvements in image acquisition and post-processing may be more critical to address. Automated approaches for segmenting the spinal cord in fMRI data would speed processing times and potentially reduce rater bias in the analysis pipeline. Future work to ensure robust processing of functional imaging data are needed to improve the sensitivity and specificity to true neural activations in the human spinal cord.

Acknowledgements

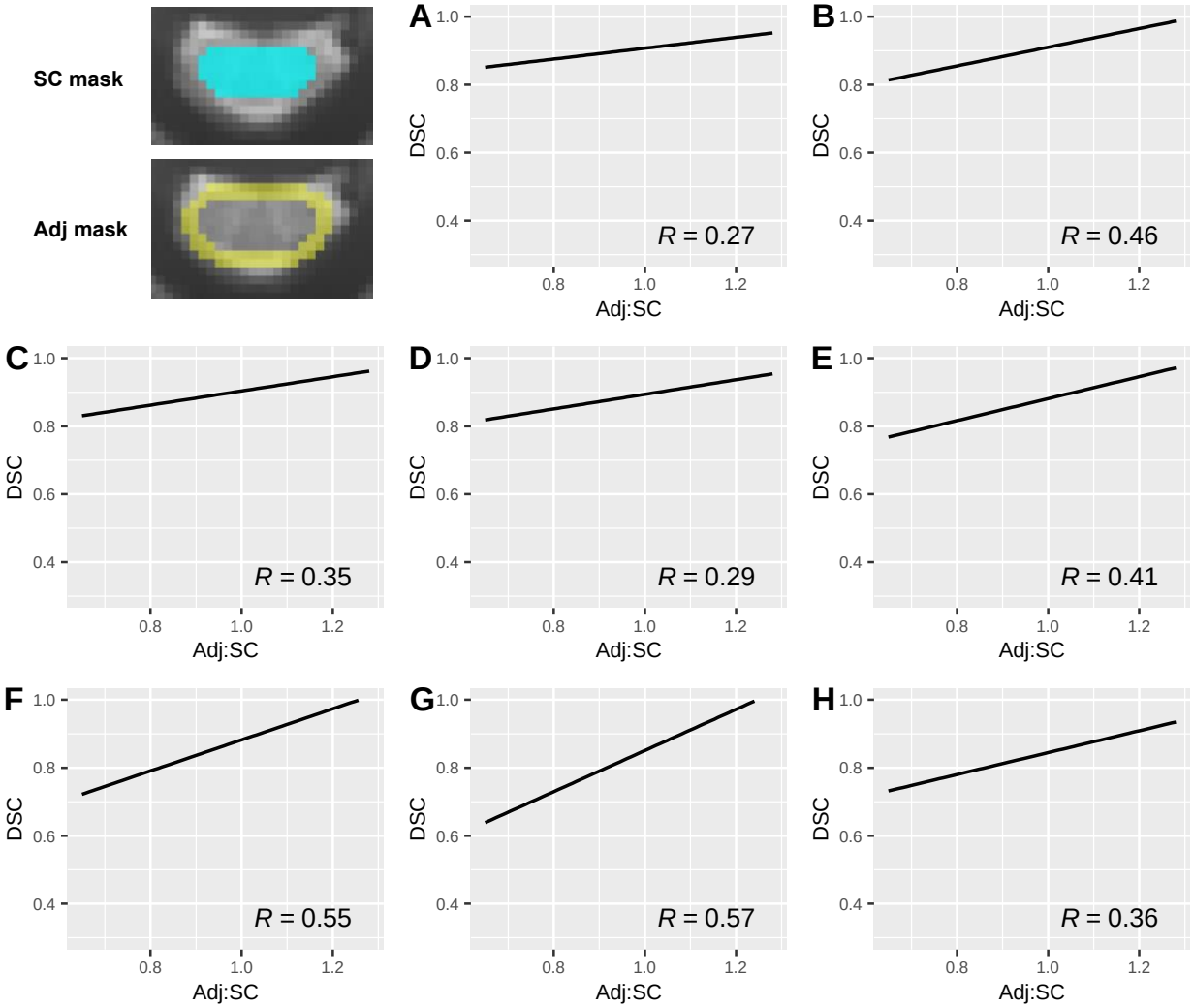
Research supported by a grant from the Craig H. Neilsen Foundation (595499), and by grants from the National Institute of Neurological Disorders and Stroke (K23NS091430, K23NS104211, and L30NS108301). K.J.H. was supported by an NIH-funded training program (T32EB025766). A.D.V. was supported by the National Science Foundation Graduate Research Fellowship (DGE-1324585) and by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health (F31NS126012). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

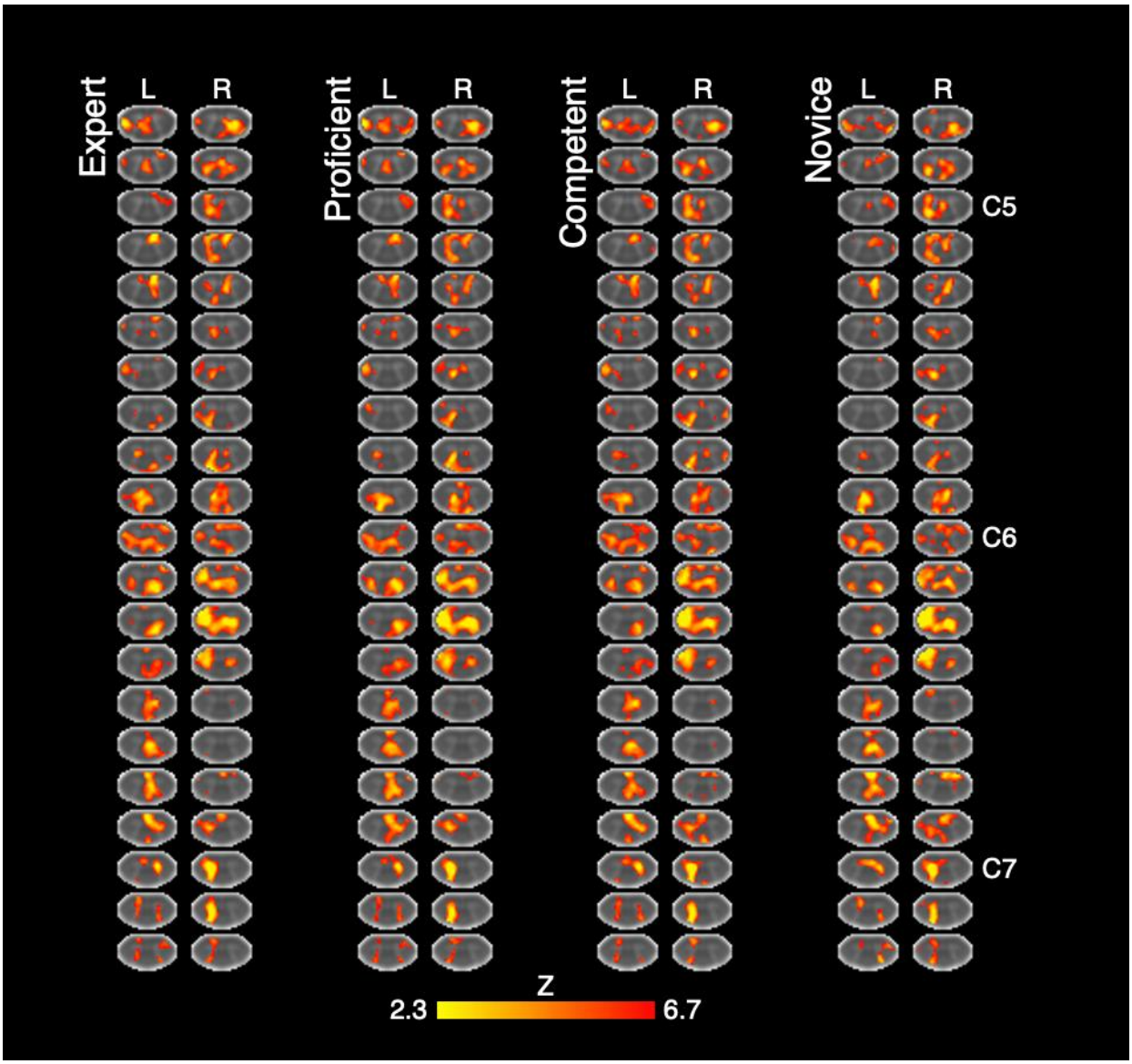
1. Landelle C, Lungu O, Vahdat S, Kavounoudias A, Marchand-Pauvert V, De Leener B, Doyon J. Investigating the human spinal sensorimotor pathways through functional magnetic resonance imaging. NeuroImage. 2021 Dec 15;245:118684.
2. Eippert F, Kong Y, Jenkinson M, Tracey I, Brooks JCW. Denoising spinal cord fMRI data: Approaches to acquisition and analysis. Neuroimage. 2017;154:255-266.doi:10.1016/j.neuroimage.2016.09.0652.
3. Kong Y, Jenkinson M, Andersson J, Tracey I, Brooks JCW. Assessment of physiological noise modelling methods for functional imaging of the spinal cord. Neuroimage.2012;60(2):1538-1549. doi:10.1016/j.neuroimage.2011.11.0773.
4. Powers JM, Ioachim G, Stroman PW. Ten key insights into the use of spinal cord fMRI. Brain sciences. 2018 Sep;8(9):173.
5. Bowring A, Nichols TE, Maumet C. Isolating the sources of pipeline-variability in group-level task-fMRI results. Human brain mapping. 2021 Aug 20.
6. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical research. 1996 Jun 1;29(3):162-73.
7. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. Fsl. Neuroimage. 2012 Aug 15;62(2):782-90.
8. Penny WD, Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE, editors. Statistical parametric mapping: the analysis of functional brain images. Elsevier; 2011 Apr 28.
9. De Leener B, Lévy S, Dupont SM, Fonov VS, Stikov N, Collins DL, Callot V, Cohen-Adad J. SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data. Neuroimage. 2017 Jan 15;145:24-43.
10. Avants BB, Tustison N, Song G. Advanced normalization tools (ANTS). Insight j. 2009 Jun 4;2(365):1-35.
11. Weber KA, Chen Y, Paliwal M, et al. Assessing the spatial distribution of cervical spinal cord activity during tactile stimulation of the upper extremity in humans with functional magnetic resonance imaging. Neuroimage. 2020;217:116905.doi:10.1016/j.neuroimage.2020.1169055.
12. Rieseberg S, Frahm J, Finsterbusch J. Two-dimensional spatially-selective RF excitation pulses in echo-planar imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine. 2002 Jun;47(6):1186-93.
13. Lichy MP, Wietek BM, Mugler III JP, Horger W, Menzel MI, Anastasiadis A, Siegmann K, Niemeyer T, Königsrainer A, Kiefer B, Schick F. Magnetic resonance imaging of the body trunk using a single-slab, 3-dimensional, T2-weighted turbo-spin-echo sequence with high sampling efficiency (SPACE) for high spatial resolution imaging: initial clinical experiences. Investigative radiology. 2005 Dec 1;40(12):754-60.
14. Pfeuffer J, Van Moortele PF De, Ugurbil K, Hu X, Glover GH. Correction of physiologically induced global off-resonance effects in dynamic echo-planar and spiral functional imaging. Magn Reson Med. 2002;47(2):344-353. doi:10.1002/mrm.100656.
15. Rieseberg S, Frahm J, Finsterbusch J. Two-dimensional spatially-selective RF excitation pulses in echo-planar imaging. Magn Reson Med. 2002;47(6):1186-1193.doi:10.1002/mrm.101577.
16. Pfeuffer J, van de Moortele PF, Yacoub E, Shmuel A, Adriany G, Andersen P, Merkle H, Garwood M, Ugurbil K, Hu X. Zoomed functional imaging in the human brain at 7 Tesla with simultaneous high spatial and high temporal resolution. Neuroimage. 2002 Sep 1;17(1):272-86.
17. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage. 2002 Oct 1;17(2):825-41.
18. Brooks JC, Beckmann CF, Miller KL, Wise RG, Porro CA, Tracey I, Jenkinson M. Physiological noise modelling for spinal functional magnetic resonance imaging studies. Neuroimage. 2008 Jan 15;39(2):680-92.

19. Glover GH, Li TQ, Ress D. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine. 2000 Jul;44(1):162-7.
20. Smith SM, Brady JM. SUSAN—a new approach to low level image processing. International journal of computer vision. 1997 May;23(1):45-78.
21. Gros C, De Leener B, Badji A, Maranzano J, Eden D, Dupont SM, Talbott J, Zhuoquiong R, Liu Y, Granberg T, Ouellette R. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. Neuroimage. 2019 Jan 1;184:901-15.
22. De Leener B, Fonov VS, Collins DL, Callot V, Stikov N, Cohen-Adad J. PAM50: Unbiased multimodal template of the brainstem and spinal cord aligned with the ICBM152 space. Neuroimage. 2018;165:170-179. doi:10.1016/j.neuroimage.2017.10.0414.
23. Verma T, Cohen-Adad J. Effect of respiration on the B0 field in the human spinal cord at 3T. Magnetic resonance in medicine. 2014 Dec;72(6):1629-36.
24. Stroman PW, Wheeler-Kingshott C, Bacon M, Schwab JM, Bosma R, Brooks J, Cadotte D, Carlstedt T, Ciccarelli O, Cohen-Adad J, Curt A. The current state-of-the-art of spinal cord imaging: methods. Neuroimage. 2014 Jan 1;84:1070-81.
25. Winklhofer S, Schoth F, Stolzmann P, Krings T, Mull M, Wiesmann M, Stracke CP. Spinal cord motion: influence of respiration and cardiac cycle. InRöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren 2014 Nov (Vol. 186, No. 11, pp. 1016-1021). © Georg Thieme Verlag KG.
26. Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage. 2001 Dec 1;14(6):1370-86.
27. Holsheimer J, Den Boer JA, Struijk JJ, Rozeboom AR. MR assessment of the normal position of the spinal cord in the spinal canal. American journal of neuroradiology. 1994 May 1;15(5):951-9.
28. Cohen-Adad J. Functional magnetic resonance imaging of the spinal cord: current status and future developments. InSeminars in Ultrasound, CT and MRI 2017 Apr 1 (Vol. 38, No. 2, pp. 176-186). WB Saunders.
29. Barry RL, Vannesjo SJ, By S, Gore JC, Smith SA. Spinal cord MRI at 7T. Neuroimage. 2018 Mar 1;168:437-51.
30. Sherman JL, Nassaux PY, Citrin CM. Measurements of the normal cervical spinal cord on MR imaging. American Journal of Neuroradiology. 1990 Mar 1;11(2):369-72.

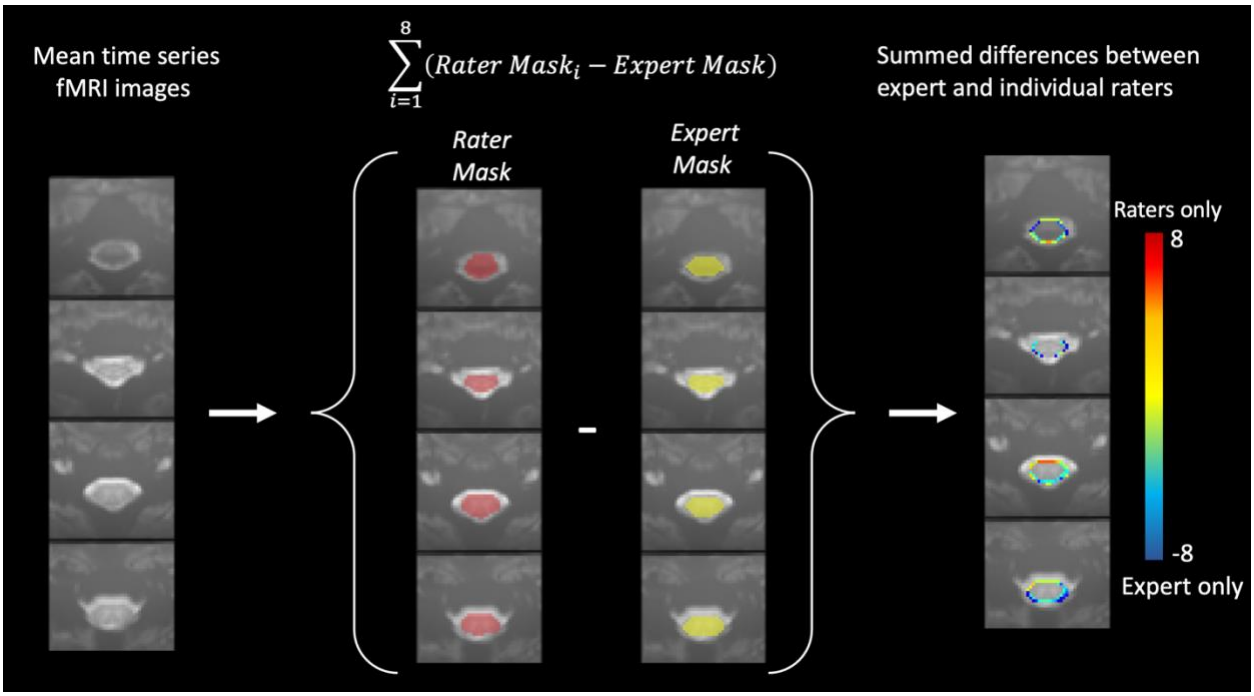
Supplementary Information



Supplemental Figure 1*: Slice-wise DSC and Adj:SC ratio correlations for every rater. Top left: a visualization of masks defining adjacent and SC voxel regions on the Expert mask. Plots A-H: DSC and Adj:SC ratio are positively correlated for every rater. Greater contrast between spinal cord and adjacent voxels on a given transverse slice is associated with increased agreement with the expert rater (positive correlation).*



Supplemental Figure 2. *Group-level activation maps from Expert, Proficient (Rater A), Competent (Rater D), and Novice (Rater H) raters. Activation is thresholded at Z-score>2.3 (cluster corrected p<0.05). Data have been transformed to the standard PAM50 template space, and the approximate level of C5-C7 spinal levels are indicated. (L=left, R=right).*



Supplemental Figure 3*: An example of the summed differences between the expert mask and the raters' mask. Eight raters and an expert contoured the spinal cord on temporal mean fMRI images. The expert mask was subtracted from each of the rater masks and the differences summed. The resultant summed difference maps are shown on the right. The difference map colorscale runs from -8 to +8, with negative values representing voxels the expert chose that were not selected by the 8 raters, and positive values being the reverse, where the raters included voxels that were not selected by the expert. The difference maps illustrate the location of disagreement between the raters and expert is at the edges of the spinal cord.*