Comparison of multi-center MRI protocols for visualizing the spinal cord gray matter

Julien Cohen-Adad[1,2,3], Eva Alonso-Ortiz[1], Stephanie Alley[1], Maria Marcella Lagana[4], Francesca Baglio[4], Signe Johanna Vannesjo[5,6], Haleh Karbasforoushan[7,8], Maryam Seif[9,10] Alan C. Seifert[11], Junqian Xu[11], Joo-Won Kim[11], René Labounek[12,13], Lubomír Vojtíšek[14], Marek Dostál[15], Jan Valošek[16], Rebecca S. Samson[17], Francesco Grussu[17,18], Marco Battiston[17], Claudia A. M. Gandini Wheeler-Kingshott[17,19,20], Marios C. Yiannakas[17], Guillaume Gilbert[21], Torben Schneider[22], Brian Johnson[23], Ferran Prados[17,18,24]

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

[2] Functional Neuroimaging Unit, CRIUGM, University of Montreal, Montreal, QC, Canada

[3] Mila - Quebec AI Institute, Montreal, QC, Canada

[4] IRCCS Fondazione Don Carlo Gnocchi ONLUS, Milan, Italy

[5] Wellcome Centre for Integrative Neuroimaging, FMRIB, University of Oxford, John Radcliffe Hospital, Oxford, UK

[6] Department of Physics, Norwegian University of Science and Technology, Trondheim, Norway

[7] Interdepartmental Neuroscience Program, Northwestern University School of Medicine, Chicago, IL, USA

[8] Department of Psychiatry and Behavioral Sciences, School of Medicine, Stanford University, Stanford, CA, USA.

[9] Spinal Cord Injury Center, Balgrist University Hospital, University of Zurich, Zurich, Switzerland

[10] Department of Neurophysics, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[11] Biomedical Engineering and Imaging Institute, Department of Radiology, Graduate School of Biomedical Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[12] Departments of Neurology and Biomedical Engineering, University Hospital Olomouc, Olomouc, Czech Republic

[13] Division of Clinical Behavioral Neuroscience, Department of Pediatrics, Masonic Institute for the Developing Brain, University of Minnesota, Minneapolis, MN, USA

[14] CEITEC - Central European Institute of Technology, Masaryk University, Brno, Czech Republic

[15] Department of Radiology and Nuclear Medicine, University Hospital Brno, Brno, Czech Republic

[16] Departments of Neurology and Biomedical Engineering, University Hospital Olomouc, Olomouc, Czech Republic

[17] Queen Square MS Centre, UCL Institute of Neurology, Faculty of Brain Sciences, University College London, London, UK

[18] Radiomics Group, Vall d'Hebron Institute of Oncology, Vall d'Hebron Barcelona Hospital Campus, Barcelona, Spain

[19] Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy

[20] Brain MRI 3T Research Centre, C. Mondino National Neurological Institute, Pavia, Italy

[21] MR Clinical Science, Philips Canada, Mississauga, ON, Canada

[22] MR Clinical Science, Philips UK, Guildford, Surrey, UK

[23] MR Clinical Development, Philips North America, Gainesville, FL, USA

[24] Universitat Oberta de Catalunya, Barcelona, Spain

Corresponding author:

Julien Cohen-Adad

Ecole Polytechnique, Pavillon Lassonde, L5610

2700, chemin de la Tour, Montréal, QC, H3T 1J4, Canada

514 340 5121 #2264

jcohen@polymtl.ca

Abstract

Purpose: Spinal cord gray matter imaging is valuable for a number of applications, but remains challenging. The purpose of this work was to compare various MRI protocols at 1.5 T, 3 T and 7 T for visualizing the gray matter.

Methods: In vivo data of the cervical spinal cord were collected from nine different imaging centers. Data processing consisted in automatically segmenting the spinal cord and its gray matter and co-registering back-to-back scans. We computed the signal-to-noise ratio using two methods (SNR_single using a single scan and SNR_diff using the difference between back-to-back scans) and the white/gray matter contrast-to-noise ratio per unit time. Synthetic phantom data were generated to evaluate the metrics performance. Experienced radiologists qualitatively scored the images. We ran the same processing on an open-access multi-center dataset of the spinal cord MRI (n = 267 participants).

Results: Qualitative assessments indicated comparable image quality for 3 T and 7 T scans. Spatial resolution was higher at higher field strength, and image quality at 1.5 T was found to be moderate to low. The proposed quantitative metrics were found to be robust to underlying changes to the SNR and contrast, however the SNR_single method lacked accuracy when there were excessive partial volume effects.

Conclusion: We propose quality assessment criteria and metrics for gray matter visualization and apply them to different protocols. The proposed criteria and metrics, the analyzed protocols, and our open-source code can serve as a benchmark for future optimization of spinal cord gray matter imaging protocols.

Keywords: MRI, acquisition, protocol, spinal cord, gray matter, image quality

Introduction

Imaging the spinal cord (SC) gray matter (GM) is useful for assessing atrophy in motor-neuron diseases such as amyotrophic lateral sclerosis (1), for studying dorsal horn atrophy in chronic pain (2), for better characterizing lesion extent in multiple sclerosis (3,4) or for improving the interpretation of SC functional MRI (5) or diffusion MRI (6–9). However, proper imaging of the SC GM is difficult due to its relatively small size and requires high spatial resolution at the expense of a lower signal-to-noise ratio (SNR) or longer acquisition times. Moreover, images are hampered by motion (e.g., swallowing, SC motion due to cerebrospinal fluid pulsation) and static susceptibility artifacts (induced by the presence of tissues with different susceptibility such as cartilage, bone, parenchyma and fat), which lead to poor fat saturation, intravoxel dephasing in gradient-recalled echo (GRE) scans and image distortions in echo-planar imaging (EPI) (10). In addition to static susceptibility effects, the B0 field varies during respiration due to the change in volume and oxygenation of inhaled air. This effect becomes more prominent with increased magnetic field strength (11,12).

The imaging protocols that are most commonly used for SC MRI and rely upon T1-weighted and T2-weighted (T1w and T2w) scans, do not provide adequate GM/white matter (WM) contrast for GM visualization and quantification. Among the preferred sequences (13) are 2D or 3D T2*-weighted (T2*w) gradient echo and 2D T1w phase sensitive inversion recovery (PSIR). In (13), the authors compared different protocols for GM imaging at 3 T based on 2D PSIR and 2D T2*w sequences across Siemens, Philips and GE vendors, providing the community with a valuable starting point for making informed decisions when it comes to GM imaging. The PSIR protocols used in that study were based on previous experience (14–16) and the 2D T2*w protocols were obtained from the 3 T cervical SC MRI spine-generic protocol (17).

The main objective of this study is to compare various imaging protocols at 1.5 T, 3 T and 7 T for visualizing GM. This article follows the *“2018 Spinal Cord Gray Matter Imaging Challenge”* that was launched at the 5th Spinal Cord MRI Workshop (http://www.spinalcordmri.org/2018/06/22/workshop.html). More specifically, this study provides (i) evaluation criteria and metrics to assess the quality of SC GM scans, (ii) an open-source and automatic analysis framework for computing those metrics, (iii) an open-access dataset from multiple centers with suggested acquisition protocols for optimal GM visualization, (iv) a comparison of those protocols using the proposed criteria and metrics and (v) a discussion about the pros/cons of various acquisition strategies.

Methods

Gray matter imaging challenge: Rules and data management

The GM imaging challenge called for MRI protocol design and pioneering data acquisition of SC images with high spatial resolution, minimal acquisition time and high GM SNR and contrast.

Protocol and data submission for the challenge was done on the Niftyweb[1] platform, with the acquisition rules listed in **Table S1**. Submission is now closed, but new participants can still run the evaluation pipeline on the challenge data or on new data using the analysis scripts (see *Inter-protocol comparison*).

To facilitate the visualization and processing of the submitted dataset, and to promote reusability of open-access material, the submitted dataset was anonymized and converted to the Brain Imaging Data Structure (BIDS) (18) and hosted on GitHub: https://github.com/sct-pipeline/gm-challenge-data. Each participant gave their consent (at the center where the data were acquired) to have their data publicly accessible.

Evaluation of imaging protocols

The comparison was divided into *quantitative* and *qualitative* assessments. All quantitative assessments were done automatically using the Spinal Cord Toolbox (SCT) (19) and custom scripts specific to this challenge (https://github.com/sct-pipeline/gm-challenge). Qualitative assessment was done by radiologists.

Quantitative assessment

Acquisition time, spatial resolution, SNR and contrast-to-noise ratio (CNR) were evaluated. Due to the difficulty in properly assessing SNR (20), we opted for two different SNR measures: one based on a single scan (SNR_single) and another based on two scans acquired back-to-back (SNR_diff). SNR and CNR were computed slice-wise and then averaged across slices.

SNR_single: While traditionally noise is estimated in the background (air), we could not do it here because (i) some images suffered from excessive ghosting in the background which would lead to overestimation of the noise standard deviation (STD) and (ii) some scans were automatically thresholded (zeroed voxels in the background) by the scanner's proprietary reconstruction pipeline. Hence we opted for computing noise in the WM to obtain a surrogate of SNR in cases where only one image was available. The WM was chosen because it pertains to the region of interest, it includes a sufficient number of voxels per slice and the signal in this region is assumed to be homogeneous slice-wise (a requirement for spatial STD computation). The steps are:

- Automatically segment the SC (21) and the GM (22) (with manual correction when needed), and compute a WM mask by subtracting the GM from the SC mask.
- A WM mask is eroded by 1 pixel (WMe) to minimize partial volume effect.
- With *S(r)* the MRI signal in voxel *r*, SNR_single is calculated as:

$$SNR\_single = \frac{mean_{r \in WMe}\{S(r)\}}{std_{r \in WMe}\{S(r)\}}$$

[1] http://niftyweb.cs.ucl.ac.uk/program.php?p=WMGM

There is no correction for the Rician distribution given that the noise is computed in a region largely above the noise floor, where the distribution is closer to a Gaussian function.

SNR_diff is the difference in SNR between two scans as in (20) and it was computed as follows:

- Volume #2 is registered to volume #1 (interpolation using nearest neighbor so as to not alter noise properties).
- With *S(r,k₁) and S(r,k₂)* the MRI signal in voxel *r* for volumes 1 and 2 respectively, SNR_diff is calculated as:

$$SNR\_diff = \frac{mean_{r \in WMe}\{S(r,k_1) + S(r,k_2)\}}{\sqrt{2} \cdot std_{r \in WMe}\{S(r,k_1) - S(r,k_2)\}}$$

CNR_single and CNR_diff were calculated by multiplying the Weber contrast by SNR_single and SNR_diff, respectively. The contrast (in percent) was computed as: 100 * |mean(WM) - mean(GM)| / mean(WM). The CNR measures were subsequently divided by the square root of the volume acquisition time (in seconds) and are called: **CNR_single/√t** and **CNR_diff/√t**.

Qualitative assessment

Two experienced radiologists scored four qualitative criteria (see **Figure 2**) for both acquisitions of each protocol. Images were presented to scorers in randomized order to minimize bias. The scoring integer scale ranged from 1 (worst) through 3 (moderate) to 5 (best). The final score for each protocol was the average of the four qualitative criteria. The median score of the two scorers was computed for each criteria as well as for the final score. The level of agreement over scorers was assessed with Spearman's rank correlation coefficient for each criteria and the final score.

Comparison with the spine-generic protocol

In order to compare the protocols submitted to the challenge with the protocol proposed for T2*w SC MRI as part of the spine-generic protocol (17), we computed SNR_single and CNR_single/√t metrics for T2*w images of the multi-subject spine-generic dataset (n = 267, all acquired at 3 T) (23). The 'diff' metrics could not be computed because the spine-generic dataset only contains a single T2*w scan for each subject. Due to slight differences in the spine-generic acquisition protocols across GE, Siemens and Philips scanners, the resulting metrics are clustered for each manufacturer.

Simulations to assess the relevance of the evaluation metrics

To assess the relevance of the proposed metrics, we generated synthetic data of the spinal cord with varying WM/GM contrasts, noise levels and smoothing factors, as done in (24). Each phantom consisted of 10 slices extracted from the PAM50 template (25) centered at the mid-C4 vertebral level. The effect of spatial resolution was assessed by smoothing the phantom with a kernel of 1 mm standard deviation. Different noise levels were then added to each phantom (additive Gaussian noise with zero mean), leading to standard deviations in the WM of 20, 5, and

1 and resulting in theoretical SNR_single levels of 10, 20, and 100. For both smoothed and unsmoothed phantoms, each simulated SNR level was modified so as to simulate different WM/GM contrast levels. This was done by fixing the signal value in WM to [100], while varying values in GM [120, 140, 160, 180], yielding contrasts of 20%, 40%, 60% and 80%. The signal in WM was fixed so that the SNR would be insensitive to the contrast (SNR was computed in the WM only). We then used these phantoms to assess the sensitivity and specificity of the evaluation criteria. We also assessed whether the measured contrast was insensitive to SNR and the other way around.

Optimal combination of echo times in multi-echo GRE acquisitions

To test whether CNR is optimized at or near T2*, we evaluated SNR, contrast, and the product of these two values (which serves as an indirect measure of CNR) in 7 T GRE images from the Mount Sinai submission (9605). The TEs varied between 3 and 19 ms, at which point localized signal drop-outs due to magnetic field inhomogeneities began to encroach on the SC and in root-sum-square combinations of these images. T2* values of 21.4 and 25.5 ms were calculated in WM and GM, respectively. Voxelwise maps of T2* had extremely high noise and, therefore, could not be accurately segmented for analysis.

The root-sum-square combination of echo images weights the contribution from each echo image by its signal intensity at each voxel, thereby maximizing SNR. However, the criteria that we intend to maximize is rather the CNR. The CNR-optimal weighting scheme would instead use the contrast or CNR of each individual echo image as the weights in a weighted sum. Four weighting schemes were evaluated: (i) the theoretical contrast ratio, calculated as the ratio of two exponential decays having time constants equal to the T2* values of WM and GM, (ii) the observed contrast in the individual echo images, (iii) the theoretical signal difference, calculated as the difference of the aforementioned exponential decays, and (iv) the observed CNR (SNR×contrast product).

Results

The results presented here can be reproduced with the following code/data versions:

- https://github.com/sct-pipeline/gm-challenge/releases/tag/v0.5
- https://github.com/sct-pipeline/gm-challenge-data/releases/tag/r20220125
- https://github.com/spine-generic/data-multi-subject/releases/tag/r20220125

Designed imaging protocols

Participating researchers designed, optimized and submitted 13 different protocols whose data were acquired over 9 MRI imaging centers. All protocols used 2D T2*w imaging, except for one which made use of a 2D T2*w scan with an additional ihMT prepulse to further suppress WM signal (Philips 9604). Two protocols were optimized for 1.5 T MRI, six for 3 T MRI and five for 7 T MRI. Each fully detailed protocol is available on the GitHub's 'gm-challenge-data' repository under each subject (file name: **sub-XXXX/anat/sub-XXXX__acq_params.pdf**).

Inter-protocol comparison

For each protocol, **Figure 1** shows a representative axial slice of an acquired image and its quantitative characteristics. The shortest scan time, highest SNR, contrast and CNR, per field strength, is shown in bold.

Figure S1 shows a pairwise comparison of both SNR methods used in this study. On average, SNR_single is 30% smaller than SNR_diff.

Final qualitative assessment scores identified that the overall image quality is highly comparable between 3 T and 7 T protocols (**Figure 2**). GM/WM contrast and sharpness were mostly higher for 7 T scans, but increased artifacts devalued their overall image quality (**Figure 2**). The image quality of 1.5 T protocols was less than moderate over most of qualitative assessments (**Figure 2**). Spearman rank correlation coefficients assessed that both scorers agreed in trends of scores over acquisitions in all qualitative assessments ($p \leq 0.009$) except the sharpness of the WM/GM border ($p = 0.114$).

Comparison with the spine-generic protocol

Figure 3 shows SNR_single and contrast measured on the T2*w images of the spine-generic multi-subject dataset. Because the gm-challenge protocols and the spine-generic protocol focused on different FOVs of different sizes, a direct and fair comparison is not fully possible. Moreover, because only one T2*w scan was acquired in the spine-generic protocol, we could not compute SNR_diff. When looking at the Siemens protocol, SNR_single and contrast of the gm-challenge results mostly overlap with the Q1-Q3 interval of the spine-generic results. The Philips gm-challenge result shows better SNR_single (15.04, above the Q1 percentile) and similar contrast (15.32, within the Q1-Q3 interval).

Validation of the quality assessment metrics

Figure 4 shows the synthetic phantoms (upper panel), and the measured contrast and SNR (lower panels). The contrast measured on the synthetic phantom showed values similar to the simulated contrast, regardless of the SNR value (**Figure 4a**, left). For the smoothed phantom, higher differences between simulated and measured contrast were obtained (**Figure 4b**, left). The measured SNR_diff was similar to the simulated SNR for each contrast (**Figure 4a**, middle), with a negligible difference for the smoothed phantom (**Figure 4b**, middle). However, the SNR_single lacked accuracy (**Figure 4a**, right) especially with smoothing (**Figure 4b**, right). This is likely due to the strong impact of partial volume effect (mixed tissue within WM mask).

Optimal combination of echo times

Figure 5 shows the results of the simulation that investigated SNR, Contrast, and pseudo-CNR as a function of echo time. As expected, for individual images, SNR decreases and contrast increases rapidly with increasing TE. The SNR×contrast product has a broad plateau between 10 and 15 ms. For root-sum-square combinations of echo images up to a given echo time (i.e., cumulative echo images), SNR is maximized at approximately 10-12 ms, while contrast increases with increasing TE. The SNR×contrast product for cumulative echoes also increases with increasing TE, but appears to plateau at approximately 17-19 ms. The plateaus in the SNR×contrast product for both individual and cumulative echo images suggests that factors besides T2* and thermal noise degrade images at TEs exceeding 15-17 ms.

All four of weighted schemes produce greater contrast than a root-sum-square combination with uniform weighting, but the root-sum-square combination with uniform weighting yields the highest CNR.

Discussion

In this article we suggest a number of criteria for evaluating spinal cord gray matter MRI and we use those criteria to assess image protocols that were submitted to the *2018 Spinal Cord Gray Matter Imaging Challenge"*. The imaging criteria, the analyzed protocols, and the open-source code which was developed for assessing image quality can serve the community as a benchmark for future protocol optimization. The following discussion expands on some of the strategies for helping the imaging community further optimize such protocols.

Evaluation criteria

One of the difficulties in organizing this challenge was to find the right balance between harmonization/simplicity (e.g., finding a set of evaluation criteria that can apply to all participants) and exhaustiveness/rigor (i.e., making sure evaluation is accurate and fair). We acknowledge there are limitations in the current design, which are discussed below.

SNR

In this study we used two different methods to compute SNR: the “diff” method, which uses the subtraction of two scans acquired back-to-back, and the “single” method, which uses a single scan where the noise variance is computed inside the WM. On average, SNR_single was 30% smaller than SNR_diff, which is likely caused by the fact that we measured the standard deviation of the signal within the WM, and not in a background region that contains pure noise. An ROI within the WM may have sources of signal variance other than noise, including partial volume effects with the cerebrospinal fluid (CSF) and the GM. From Dietrich et al. (20), SNR_diff is closer to the “true” SNR (ie: the “mult” or “nema” approach), which is also confirmed by the simulation results (**Figure 4**). So we considered the “diff” results from the present study to be more reliable. The “single” method has the advantage of being computed with only one scan, hence we were able to compute SNR from a retrospective database of 267 individuals from the spine-generic project.

SNR is directly related to the average of the magnitude image in the region of interest; in our case, the WM. Therefore, if a sequence yields low signal in the WM, the SNR will consequently be low (assuming constant noise variance). For example, let’s consider two datasets (A and B) with the same noise amplitude everywhere in the image, the same mean signal in the GM, but the mean signal in the WM being lower in dataset A. The SNR calculated in the GM would be the same in datasets A and B, but the SNR calculated in the WM would be lower in dataset A while the WM-GM contrast would be higher in dataset A. The contrast on the other hand will be increased by a low value in the denominator. This is observed in the Oxford (9611Ses2) submission, which shows a relatively low SNR value in the WM, but high contrast. If SNR was measured in another region, the apparent relative performance across protocols would likely differ.

Another (related) consideration is that T2* is driven by the field strength and the orientation of myelinated fibers (26). So, it is not surprising that some of the 7 T scans show a relatively lower SNR compared to 1.5 and 3 T scans, even though higher field strength should *in principle* yield higher SNR. Moreover, to compare SNR between field strengths one should also account for voxel volume and acquisition time. An SNR efficiency measure that corrects for those would be interesting to include in the future.

When the scanner saves its "magnitude" data, it may already be slightly filtered (e.g., using a Fermi filter to reduce ringing), which would change the inherent noise profile before the SNR is calculated. Also, the use of a multi-channel coil induces spatially variant noise properties, hence there is a bias when computing noise STD across space, as was done here. Other methods exist that are more accurate than the ones used here. For example, acquiring two scans back-to-back, one with and the other without transmit voltage, to estimate noise STD without any bias from coil combination (20). This method requires collecting and processing raw data, which was not done for the sake of simplicity.

Contrast and CNR

One of the difficulties in estimating contrast is obtaining a reliable measure of the average signal within each region, in this case WM and GM. In order to minimize partial volume effects, we

eroded the WM mask by 1 voxel. We decided not to do the erosion for the GM mask because this would have resulted in a very low number of remaining voxels, and hence low statistical power. If we had access to partial volume information, we could have used Gaussian mixture modeling to account for partial volume effect at the CSF/WM/GM interfaces, as was done in (24). Such information could be derived from a high resolution atlas registered to each dataset, and then downsampled at the native resolution of the data. This was not done here because such registration is critical, and any mis-registration would yield other errors which we preferred not to address within the scope of this study. Contrast is also influenced by slice orientation. This is mostly due to partial volume effects, however, B_0 inhomogeneity and susceptibility differences between discs, bones and air degrade the contrast in gradient-echo (GRE) based sequences as well.

A study by Papinutto et al. (Papinutto and Henry 2019) reported an average CNR(GM/WM) of 1.56 on Siemens 3T datasets. To be able to compare this value with our results, we computed the CNR_single without normalizing by the square root of the acquisition time and without converting it into percent value. We considered only the Siemens 3T results, yielding a CNR_single(GM/WM) of 2.81 +/- 0.56 (mean +/- SD). This is slightly higher than the average value reported by Papinutto et al.

Resolution

The spatial resolution impacts the 'sharpness' of an image, or our ability to distinguish between two small objects. A measure of sharpness can be obtained by computing the laplacian of the image, then computing the mean of the Laplacian inside the SC. However, this measure is also sensitive to the noise level: the higher the noise, the higher the Laplacian. For this reason we only considered the acquired spatial resolution (field of view divided by matrix size), although we should keep in mind that the effective resolution is also affected by the use of partial Fourier and additional filtering done by each manufacturer, even though one criteria of the challenge was specifically to not add reconstruction filters (e.g., hanning windowing).

Choice of sequence parameters

Below are some useful considerations when optimizing SC GM imaging. More details are given in the spine-generic protocol study (17).

2D vs. 3D

Compared to 3D imaging, multislice (2D) imaging is more robust to subject motion (if the subject moves, this will not affect the entire image), has no aliasing at the edges, and there are no issues with the B1+ profile (3D images have imperfect slab profiles creating lower flip angles at the edges, which requires one to discard 2-3 slices at the edge). On the other hand, 3D acquisitions are more SNR efficient.

Phase encoding direction

Because motion is predominantly along the A-P direction, when possible, it would be preferable to phase-encode along the R-L direction. However, when imaging below the cervical cord, this becomes difficult because the shoulders and arms will alias onto the image.

Saturation band

The traditional purpose of saturation bands is to suppress unwanted signals, in order to avoid wrap-around artifacts. Because these spatial saturation pulses are usually transmitted at a different carrier frequency, they produce a slight magnetization transfer (MT) effect, which in turn alters WM/GM contrast. Therefore, they could be used to enhance WM/GM contrast, assuming that the MT effect suppresses signal from WM more than from GM, and that the main contrast is T2*-like (i.e. brighter GM).

Optimal combination of echo times

The majority of the submitted protocols relied on T2*w imaging with multiple echo times. In T2*w image acquisitions, knowledge of the T2* relaxation times of the two tissue types whose contrast is to be optimized can aid in the creation of an imaging protocol. While SNR is highest at the shortest echo time (TE), T2* contrast increases with increasing TE. However, in practice, neither SNR nor contrast should be optimized in isolation. Instead, efforts should be made to optimize the CNR or CNR per unit time. Under a simplistic assumption of pure thermal noise, CNR was shown to be optimized at TE = T2* (27). Other factors, such as magnetic field inhomogeneity and limitations on total scan time, may favor shorter TE, as does the increased physiological noise at higher TE (12). The latter factor may explain why the root-sum-square echo combination, which upweights early echoes, was here observed to have higher CNR than contrast-weighted echo combinations, which should theoretically be optimal under pure thermal noise.

An additional consideration in multi-echo GRE sequences is the choice of monopolar versus bipolar readout. Bipolar readouts allow for TEs to be spaced more closely, yielding increased SNR, but may result in different patterns of spatial distortion between even and odd echoes (positive and negative readouts) due to background magnetic field inhomogeneities[2]. The mis-registration between the even and odd echoes would introduce blurring when combining all echoes. Monopolar readouts produce a set of echoes with compatible patterns of spatial distortion at some cost to SNR and CNR.

Acknowledgements

We thank Alexandru Foias and Nicolas Pinon for helping with the generation of figures, and Pavla Hanzlíková (from the Department of Radiology, University of Ostrava, Czechia) for helping with the qualitative assessment. This study was funded by the Funded by the Canada Research Chair

[2] https://raw.githubusercontent.com/sct-pipeline/gm-challenge/master/doc/fig_monopolar_bipolar.gif

in Quantitative Magnetic Resonance Imaging [950-230815], the Canadian Institute of Health Research [CIHR FDN-143263], the Canada Foundation for Innovation [32454, 34824], the Fonds de Recherche du Québec - Santé [28826], the Natural Sciences and Engineering Research Council of Canada [RGPIN-2019-07244], the Canada First Research Excellence Fund (IVADO and TransMedTech), the Courtois NeuroMod project, the Quebec BioImaging Network [5886, 35450], the Czech Health Research Council [NV18-04-00159], the Ministry of Education Youth and Sport of the Czech Republic [LM2015062, Czech-BioImaging project] and the Ministry of Health of the Czech Republic [65269705, project for conceptual development in research organizations], EU Horizon 2020 (CDS-QuaMRI 634541), Engineering and Physical Sciences Research Council (EPSRC EP/R006032/1 and EP/I027084/1), INSPIRED (Spinal Research, UK; Wings for Life, Austria; Craig H. Neilsen Foundation, USA), UK Multiple Sclerosis Society (grants 892/08 and 77/2017), Department of Health's NIHR BRC (R&D 03/10/RAG0449), Guarantors of Brain post-doctoral non-clinical fellowships, the US National Institute of Neurological Disorders and Stroke (NIH/NINDS K01-NS105160), Beatriu de Pinós postdoctoral fellowships (2020 BP 00117, Secretary of Universities and Research, Government of Catalonia).

Bibliography

1. Paquin M-Ê, El Mendili MM, Gros C, Dupont SM, Cohen-Adad J, Pradat P-F. Spinal Cord Gray Matter Atrophy in Amyotrophic Lateral Sclerosis. AJNR Am. J. Neuroradiol. 2018;39:184–192.

2. Jutzeler CR, Huber E, Callaghan MF, et al. Association of pain and CNS structural changes after spinal cord injury. Scientific Reports 2016;6 doi: 10.1038/srep18534.

3. Calabrese M, Favaretto A, Martini V, Gallo P. Grey matter lesions in MS: from histology to clinical implications. Prion 2013;7:20–27.

4. Agosta F, Pagani E, Caputo D, Filippi M. Associations between cervical cord gray matter damage and disability in patients with multiple sclerosis. Arch. Neurol. 2007;64:1302–1305.

5. Kornelsen J, Mackey S. Potential clinical applications for spinal functional MRI. Curr. Pain Headache Rep. 2007;11:165–170.

6. Wheeler-Kingshott CA, Stroman PW, Schwab JM, et al. The current state-of-the-art of spinal cord imaging: applications. Neuroimage 2014;84:1082–1093.

7. By S, Smith AK, Dethrage LM, et al. Quantifying the impact of underlying measurement error on cervical spinal cord diffusion tensor imaging at 3T. J. Magn. Reson. Imaging 2016;44:1608–1618.

8. Massire A, Rasoanandrianina H, Taso M, et al. Feasibility of single-shot multi-level multi-angle diffusion tensor imaging of the human cervical spinal cord at 7T. Magn. Reson. Med. 2018;80:947–957.

9. Labounek R, Valošek J, Horák T, et al. HARDI-ZOOMit protocol improves specificity to microstructural changes in presymptomatic myelopathy. Sci. Rep. 2020;10:17529.

10. Cohen-Adad J, Wheeler-Kingshott C. Quantitative MRI of the Spinal Cord. Academic Press; 2014.

11. Verma T, Cohen-Adad J. Effect of respiration on the B0 field in the human spinal cord at 3T. Magn. Reson. Med. 2014;72:1629–1636.

12. Vannesjo SJ, Miller KL, Clare S, Tracey I. Spatiotemporal characterization of breathing-induced B0 field fluctuations in the cervical spinal cord at 7T. Neuroimage 2018;167:191–202.

13. Papinutto N, Henry RG. Evaluation of Intra- and Interscanner Reliability of MRI Protocols for Spinal Cord Gray Matter and Total Cross-Sectional Area Measurements. J. Magn. Reson. Imaging 2019;49:1078–1090.

14. Papinutto N, Schlaeger R, Panara V, et al. Age, Gender and Normalization Covariates for Spinal Cord Gray Matter and Total Cross-Sectional Areas at Cervical and Thoracic Levels: A 2D Phase Sensitive Inversion Recovery Imaging Study. PLOS ONE 2015;10:e0118576 doi: 10.1371/journal.pone.0118576.

15. Papinutto N, Datta E, Zhu AH, et al. Multisite feasibility study of spinal cord gray matter and total cord areas measurements on 2D phase sensitive inversion recovery images. In: Proc 24th

Annual Meeting ISMRM, Singapore. ; 2016.

16. Papinutto N, Schlaeger R, Panara V, et al. 2D phase-sensitive inversion recovery imaging to measure in vivo spinal cord gray and white matter areas in clinically feasible acquisition times. Journal of Magnetic Resonance Imaging 2015;42:698–708 doi: 10.1002/jmri.24819.

17. Cohen-Adad J, Alonso-Ortiz E, Abramovic M, et al. Generic acquisition protocol for quantitative MRI of the spinal cord. Nat. Protoc. 2021;16:4611–4632.

18. Gorgolewski KJ, Auer T, Calhoun VD, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data 2016;3:160044.

19. De Leener B, Lévy S, Dupont SM, et al. SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data. Neuroimage 2017;145:24–43.

20. Dietrich O, Raya JG, Reeder SB, Reiser MF, Schoenberg SO. Measurement of signal-to-noise ratios in MR images: influence of multichannel coils, parallel imaging, and reconstruction filters. J. Magn. Reson. Imaging 2007;26:375–385.

21. Gros C, De Leener B, Badji A, et al. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. Neuroimage 2019;184:901–915.

22. Perone CS, Calabrese E, Cohen-Adad J. Spinal cord gray matter segmentation using deep dilated convolutions. Sci. Rep. 2018;8:5966.

23. Cohen-Adad J, Alonso-Ortiz E, Abramovic M, et al. Open-access quantitative MRI data of the spinal cord and reproducibility across participants, sites and manufacturers. Sci Data 2021;8:219.

24. Levy S, Benhamou M, Naaman C, Rainville P, Callot V, Cohen-Adad J. White matter atlas of the human spinal cord with estimation of partial volume effect. Neuroimage 2015;119:262–271.

25. 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.

26. Cohen-Adad J, Polimeni JR, Helmer KG, et al. T2* mapping and B0 orientation-dependence at 7T reveal cyto- and myeloarchitecture organization of the human cortex. Neuroimage 2012;60:1006–1014.

27. Van de Moortele P-F, Ugurbil K, Lehericy S. Is T2* always the optimum Echo Time in BOLD fMRI? Challenging a common concept with a new Contrast to Noise Ratio BOLD model. In: Proceedings of the 16th Annual Meeting of ISMRM, Toronto, Canada. ; 2008.

Figure legends

Figure 1. Representative images for each protocol with its quantitative assessment. Protocols are ordered by field strength: 1.5 T (green), 3 T (black) and 7 T (red) and by submission ID (in brackets, next to the center). CNRs are expressed in percent. The best CNR per unit time, per field strength, is indicated with bold font. Each image corresponds to an axial slice centered at the C2/C3 intervertebral disc. Resolution is in mm. "Echoes / nav" corresponds to the number of echoes and the number of averages (combined with root sum squared except for site 'Philips' where all echoes were summed). Additional quantitative metrics (Contrast, CNR) can be downloaded from GitHub[3].

Figure 2. Qualitative assessment of MRI protocols. The top plot indicates the final scores for the qualitative assessment, which are taken to be the average of the four qualitative criteria shown in the remaining four plots. The y-axis is the integer score of the scale from 1 (worst) through 3 (moderate) to 5 (best). For the criteria 'Signal drop-out due to intravoxel dephasing', a low score means "strong signal drop-out" (i.e., less signal). Each criteria was assessed by two independent scorers whose scores are indicated with unique markers at left side for test and at right side for retest scans around the cyan-line median of all scores per data submission (i.e. 4 scores per submission). The value "r" represents the Spearman rank correlation coefficient assessing a level of agreement between two scorers. The value "p_r" represents the p-value of the correlation coefficient.

Figure 3. SNR_single (left) and CNR_single/√t (right) computed on the T2*-weighted data from the multi-subject dataset of the spine-generic project (n=267, all acquired at 3 T). Each panel shows the individual data (plot), the median and quartiles (box plots), the mean (triangle) and the distribution (violin plot). Outliers (diamonds) are defined as being outside the 1.5xIQR (IQR: interquartile range).

Figure 4. Interplay between evaluation metrics. These simulations are based on phantoms constructed with various levels of contrast: 20%, 40%, 60% and 80% and SNR: 5 (blue), 20 (yellow) and 100 (red). WM and GM masks derived from the PAM50 template (25) are thresholded at 0.9 to be used with the weighted average method to extract signal in the WM and GM respectively. Results show the evaluation metrics: "Measured Contrast", "Measured SNR_diff" and "Measured SNR_single" as a function of the simulated contrast and SNR, without (a) and with 1 mm kernel smoothing (b).

Figure 5. SNR, Contrast, and pseudo-CNR (SNR*Contrast) as a function of echo time. Here, (a) SNR_single, WM-GM contrast, and their product (an indirect metric of contrast-to-noise ratio, which we call pseudo-CNR), are plotted against echo time for individual images at given echo times (blue), for root-sum-square combinations of echo images up to a given echo time ('cumulative', green), and for root-sum-square combinations of echo images beginning with a given echo time ('anti-cumulative', red). A montage (b) of the underlying individual, cumulative,

[3] https://github.com/sct-pipeline/gm-challenge/releases/download/v0.5/gmchallenge_20220114_204833.zip

and anti-cumulative echo images illustrates the changes in SNR and contrast as echoes are added. An additional montage (c) of weighted echo combinations and their SNR, contrast, and pseudo-CNR (SNR*Contrast) is also shown.

Supporting Information

Item	Value	Comment
Field strength	1.5 T, 3 T or 7 T	
Coils	Product or custom	If researchers have a custom coil, they are encouraged to send data acquired with custom and product coils for comparison.
Sequence	Product or custom	If product, please indicate if a license is required. If custom, please indicate availability (e.g. WIP, C2P).
Acquisition time	10 min max	As fast as possible (speed is an evaluation criteria).
Slice thickness	3mm or less(*)	There is a tradeoff between thick slices (cons: intravoxel dephasing in T2*w) and thin slices (cons: lower SNR).
FOV	Centered on the C2-C3 disc with 50mm coverage in S-I direction	There is no restriction with respect to the slice gap (i.e. does not need to be contiguous).
Interpolation	None	Most scanners have an automatic k-space zero-padding that must be unchecked.
Filter	None	Raw and elliptical filters should not be used as these affect noise properties.
Type of data	Calculated map (e.g. T2 map) or raw data (e.g. T2w) accepted	The method to calculate the map should be reported.
Number of acquisitions	2	Two scans with the exact same parameters, without repositioning, need to be submitted in order to compute SNR using the 'diff' method.

Table S1. Acquisition criteria for submitting data to the challenge. Acronyms: Signal-to-noise ratio (SNR), T2-weighted (T2w), T2*-weighted (T2*w), field of view (FOV), superior-inferior (S-I), difference (diff). (*): The submission "Juntendo (9669)" used 5 mm slices.

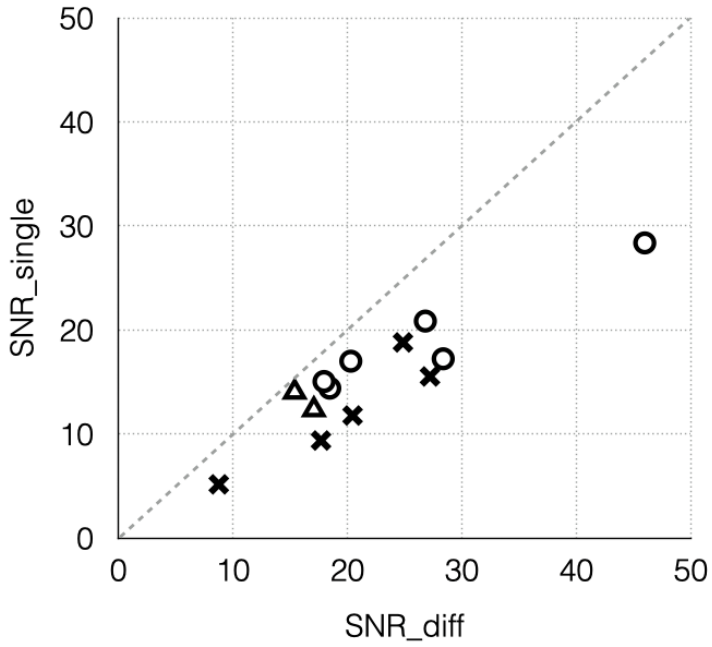

Figure S1. Pairwise comparison of the two SNR methods used in this study, showing data acquired at 1.5 T (triangle), 3 T (circle) and 7T (cross). The SNR_diff method uses the subtraction of two scans acquired back-to-back, and the SNR_single method uses a single scan where the noise variance is computed inside the WM. The dashed line corresponds to no difference between the two methods.