NeuroImage 209 (2020) 116489

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Generalised boundary shift integral for longitudinal assessment of spinal cord atrophy

Ferran Prados [a,b,c,*,1], Marcello Moccia [b,d,1], Aubrey Johnson [e], Marios Yiannakas [b], Francesco Grussu [b,f], Manuel Jorge Cardoso [g], Olga Ciccarelli [b], Sebastien Ourselin [g], Frederik Barkhof [a,b,h,2], Claudia Wheeler-Kingshott [b,i,j,2]

[a] Centre for Medical Image Computing (CMIC), Department of Medical Physics and Bioengineering, University College London, 90 High Holborn, London, WC1V 6LJ, UK
[b] NMR Research Unit, Queen Square MS Centre, Department of Neuroinflammation, UCL Queen Square Institute of Neurology, Faculty of Brain Sciences, University College London, Russell Square, London, WC1B 5EH, UK
[c] e-health Center, Universitat Oberta de Catalunya, Barcelona, Spain
[d] Multiple Sclerosis Clinical Care and Research Centre, Department of Neurosciences, Federico II University, Naples, Italy
[e] Smith College, Northampton, MA, USA
[f] Centre for Medical Image Computing (CMIC), Department of Computer Science, University College London, 90 High Holborn, London, WC1V 6LJ, UK
[g] Department of Biomedical Engineering & Imaging Sciences, King's College London, UK
[h] Dept. of Radiology & Nuclear Medicine, VU University Medical Centre, Amsterdam, the Netherlands
[i] Brain MRI 3T , UKCenter, IRCCS Mondino Foundation, Pavia, Italy
[j] Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy

A B S T R A C T

Spinal cord atrophy measurements obtained from structural magnetic resonance imaging (MRI) are associated with disability in many neurological diseases and serve as *in vivo* biomarkers of neurodegeneration. Longitudinal spinal cord atrophy rate is commonly determined from the numerical difference between two volumes (based on 3D surface fitting) or two cross-sectional areas (CSA, based on 2D edge detection) obtained at different time-points. Being an indirect measure, atrophy rates are susceptible to variable segmentation errors at the edge of the spinal cord. To overcome those limitations, we developed a new registration-based pipeline that measures atrophy rates directly. We based our approach on the generalised boundary shift integral (GBSI) method, which registers 2 scans and uses a probabilistic XOR mask over the edge of the spinal cord, thereby measuring atrophy more accurately than segmentation-based techniques. Using a large cohort of longitudinal spinal cord images (610 subjects with multiple sclerosis from a multi-centre trial and 52 healthy controls), we demonstrated that GBSI is a sensitive, quantitative and objective measure of longitudinal spinal cord volume change. The GBSI pipeline is repeatable, reproducible, and provides more precise measurements of longitudinal spinal cord atrophy than segmentation-based methods in longitudinal spinal cord atrophy studies.

1. Introduction

Spinal cord atrophy is a measure of overall spinal cord damage and has important clinical correlates in a number of neurological diseases. In multiple sclerosis (MS), spinal cord atrophy occurs from the early phases of the disease and is associated with overall disability and clinical progression (Ciccarelli et al., 2019; Moccia, Ruggieri, et al., 2019). Similarly, in amyotrophic lateral sclerosis (ALS), spinal cord atrophy predicts disease progression, respiratory failure and survival rate (El Mendili et al., 2019). Furthermore, spinal cord atrophy can occur as a consequence of spinal cord injury and its monitoring over time can shed light on the most aggressive aspects of the disease (Denecke et al., 2019). As such, spinal cord atrophy has been suggested as a possible endpoint in studies of neuroprotection (Antonescu et al., 2018; Moccia, Ruggieri, et al., 2019).

Different MS clinical trials have already used spinal cord atrophy as a secondary outcome measure (Kalkers et al., 2002; Leary et al., 2003; Lin et al., 2003; Montalban et al., 2009; Kapoor et al., 2010; Yaldizli et al., 2015; Tur et al., 2018), but yielded inconclusive or negative results. Those disappointing results may be, at least in part, due to the relatively high measurement noise and low reproducibility of the segmentation-based methods (Prados and Barkhof, 2018; Moccia et al., 2017). Currently, spinal cord atrophy is determined by numerical subtraction of volume (based on 3D surface fitting) or cross-sectional area (CSA) (based on 2D edge detection on serial images) obtained separately

* Corresponding author. 1st Floor, 90 High Holborn, London, WC1V 6LJ, UK.
E-mail address: f.carrasco@ucl.ac.uk (F. Prados).
[1] Authors contributed equally.
[2] Senior authors contributed equally.

https://doi.org/10.1016/j.neuroimage.2019.116489
Received 11 October 2019; Received in revised form 18 December 2019; Accepted 21 December 2019
Available online xxxx

at each time-point, providing what could be considered an *indirect* estimates of atrophy rates (Wheeler-Kingshott et al., 2014; Stroman et al., 2014). This strategy could introduce noise due to inconsistency in segmenting the exact same region for two different time-points. On the contrary, brain atrophy measures have been a cornerstone in the study of interventions with putative neuroprotective effects (Montalban et al., 2017; Kappos et al., 2018; Tur et al., 2018), because of the application of registration-based methods that provide *direct* estimates of brain atrophy, such as the Structural Image Evaluation using Normalization of Atrophy (SIENA) (Smith et al. 2000, 2001) and the Boundary Shift Integral (BSI) method (Freeborough and Fox, 1997; Leung et al. 2010, 2012; Prados et al., 2015). Both SIENA and BSI have reduced sample size requirements to detect significant differences between groups or over time, and are nowadays well-established methods to measure longitudinal brain atrophy in clinical trials and in observational studies for neurodegenerative diseases (Altmann et al., 2009; Schott et al., 2010).

Here, we present a specific pipeline based on the generalised formulation of BSI for the quantification of spinal cord atrophy using direct estimates. Possible consequences for the design of clinical trials and observational studies (e.g., sample size) are evaluated as a benchmark between techniques.

2. Material and methods

2.1. Pipeline overview

A graphic overview of the pipeline is presented in Fig. 1 and is applicable to datasets with T1-weighted (T1-w) sequences with identical parameters, ideally using 1 mm isometric voxel and with acquisitions at two time-points for each subject. The first step is the manual or automatic segmentation of the spinal cord from T1-w images (Prados et al., 2016; Yiannakas et al., 2016; Yiannakas et al., 2012; Horsfield et al., 2010; Gros et al., 2019; De Leener, Lévy, et al., 2017). Afterwards, the extracted masks are used to compute a ring surrounding the spinal cord to scale the signal intensity of the images accounting for the presence of the noise floor (Jones and Basser, 2004); for this step the signal intensities in the whole 3D volume are corrected using a fast version of the adaptive non-local means filter algorithm (Tristán-Vega et al., 2012). Then, an intensity inhomogeneity correction is applied to the 3D data using the N4 algorithm (Tustison et al., 2010). Once images are corrected for noise and intensity non-uniformities, both spinal cord time-points are straightened using a specific software available within the spinal cord toolbox (SCT) (De Leener, Mangeat, et al., 2017). This is an essential step to facilitate the registration between baseline and follow-up scans, as it removes the difference in curvature between time-points due to subject positioning in the scanner. Both spinal cords are then registered to the half-way space using a symmetric, affine and inverse-consistent method (Modat et al., 2014). To reduce the residual bias field and homogenise the grey scale between both registered time-points, a symmetric differential bias correction is applied (Lewis and Fox, 2004). Finally, using the generalised boundary shift integral (GBSI) (Prados et al., 2015), we obtain atrophy estimates between the two time-points. Details of specific steps are provided in the following sections.

2.1.1. Spinal cord segmentation

The whole cord is segmented (i.e., white and grey matter together), defining the spinal cord boundaries and the cranio-caudal extension of tissue over which the GBSI estimates are required. This segmentation is computed separately and independently for each time-point, over the same spinal cord segments. In this study, we used the spinal cord segment C2–C5, putting the landmarks in the middle of the corresponding spinal cord disks, but other sections could be equally used. The segmentation can be obtained manually, semi- or fully-automatically using a wide range of techniques, such as the active surface method available in JIM (JIM 6.0, Xinapse Systems, Aldwincle, UK) (Horsfield et al., 2010), tools from the SCT (De Leener, Lévy, et al., 2017; Gros et al., 2019; Yiannakas et al., 2016), or other available techniques (Prados et al., 2016). The extracted spinal cord segmentation can be represented with a hard (binary) or a soft (probability) mask. As the acquired spinal cord extension can vary by a few slices between time-points, e.g. due to positioning of the subject in the scanner, the longitudinal atrophy is computed over the intersection of these regions of interest, discarding the pixels that are not covered at both time-points.

In this paper, percentage spinal cord volume change (PCVC) was obtained as the difference between follow-up and baseline CSA, divided by baseline CSA and multiplied by 100. Segmentation masks for computing CSA were then used as inputs for the BSI pipeline.

2.1.2. Image denoising

The original T1-w images are denoised using a fast version (Tristán-Vega et al., 2012) of the adaptive non-local mean filter (Buades et al., 2005) using the mask from the segmented spinal cord. In detail, for computing the root power of the noise, we calculate the standard deviation of the signal in a ring within the cerebrospinal fluid (CSF) and scale it to account for the presence of a noise floor (Jones and Basser, 2004) (this approach is standard on T1-w images, where this can be obtained from regions where signal from CSF is suppressed). The ring within the CSF is derived by dilating the spinal cord mask obtained as described

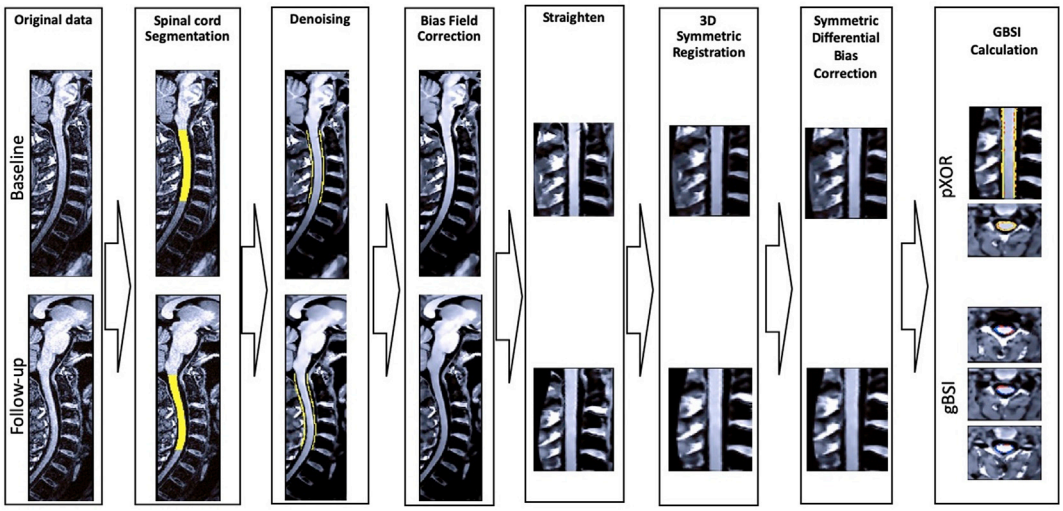

Fig. 1. Spinal cord longitudinal atrophy computation pipeline using GBSI (Generalised Boundary Shift Integral).

above (Section 2.1.1) with 2 pixel layers and then subtracting the mask after dilating the spinal cord mask only once. Prior to calculating the standard deviation, we discard any value of voxels within the extracted ring that are more than 2 standard deviations above the mean, in order to discard intensity values from nerve roots or other spurious signal intensities.

2.1.3. Inhomogeneity correction

Data are corrected for intensity inhomogeneity using N4 (Tustison et al., 2010) only over the region determined by the two times dilated spinal cord mask. The following parameters are used (Prados et al., 2015): full width at half maximum (FWHM) = 0.05, convergence threshold = 0.0001 and a maximum number of iterations = 1000.

2.1.4. Cord straightening

A common challenge in longitudinal spinal cord studies is the variability of the position of the subject within the MRI scanner between time-points. To remove any difference in the resulting cord curvature between time-points, a robust and accurate method for straightening magnetic resonance images of the spinal cord is used, based on the previously computed spinal cord segmentation (De Leener, Mangeat, et al., 2017). The main feature of this method is that it preserves spinal cord topology, which is essential for measuring subtle changes in spinal cord edges when using GBSI. The straightening method is freely available as part of the SCT software package (De Leener, Lévy, et al., 2017).

2.1.5. Half-way space registration

After straightening both time-points, images are registered to the half-way space using an affine transformation in order to avoid biases that would be introduced if registering one time-point to the other (Smith et al., 2000; Reuter et al., 2012; Leung et al., 2012). This step is achieved using an inverse-consistent and symmetric algorithm (Modat et al., 2014). Once the transformations are obtained, images and corresponding masks for each time-point are linearly resampled to the common half-way space.

2.1.6. Differential bias correction

A longitudinal differential bias correction method is used to remove the residual intensity inhomogeneity-derived differences between the baseline and the repeated images (radius = 5 voxels) (Lewis and Fox, 2004). This step is needed, despite the previous cross-sectional inhomogeneity correction, to avoid artificial atrophy values by the remaining intensity differences from the cross-sectional inhomogeneity correction method.

2.1.7. Intensity normalization

Another image pre-processing step, i.e. prior to computing the GBSI, is the normalization of the image intensities (Leung et al., 2010) and the extraction of the probabilistic area from which GBSI will be computed (Prados et al., 2015). The intensity normalization of the baseline and follow-up images is obtained from a linear regression between the average tissue intensity inside the cord and inside the CSF. The tissue intensity values are computed using a k-means algorithm (k = 2) which is delimitated by a region of interest obtained from each input mask (after 2 dilations).

2.1.8. Probabilistic XOR

The probabilistic XOR mask is obtained from the half-way linearly-resampled segmentation masks following the same steps already introduced for the brain GBSI calculations (Prados et al., 2015). This mask identifies the voxels with high probability to be tissue at the edge of the cord.

2.1.9. Atrophy computation

Finally, the GBSI is computed on a voxel-by-voxel basis as the difference in intensity between the baseline and the follow-up image within a clipped window that can be fixed (Freeborough and Fox, 1997) or adaptive (Leung et al., 2010) and can be obtained from the two k-means class values. The clipped window goal is to catch the difference between tissue intensities at the two time-points, reducing the background influence. Then the intensity differences are weighted by the probabilistic XOR mask voxel-wise. For the spinal cord GBSI, we used a predetermined clipping window. To increase robustness, the "forward" and "backward" BSI (Leung et al., 2010) is calculated for each pair of images (i.e. swapping baseline and follow-up images and repeating the intensity normalization, probabilistic XOR and atrophy computation steps), and the mean of the results is included.

PCVC was calculated by dividing the GBSI value by the binarized, straightened and registered baseline cord mask volume.

2.2. Software

The N4, denoising, differential bias correction and GBSI methods are all freely available as part of NifTK package at https://cmiclab.cs.ucl.ac.uk/CMIC/NifTK. For registration purposes, we used NiftyReg software package. SCT has been used for straightening, and can be found at https://github.com/neuropoly/spinalcordtoolbox. GBSI has been implemented in this paper using these software packages, however, other software packages with the same goal could be used in each step.

2.3. MRI data

For this study, we used three different MRI datasets to assess the performance of GBSI versus CSA. First, we used a retrospective single-centre test dataset with healthy controls only (Yiannakas et al., 2016) to compare the reproducibility of measuring the absence of longitudinal spinal cord atrophy (PCVC) with GBSI using three different CSA segmentation techniques (JIM, SCT Propseg and SCT DeepSeg) (Yiannakas et al., 2016). Secondly, we computed PCVC with GBSI and CSA using JIM over T1-w data from a large multi-site clinical trial in primary progressive MS (Lublin et al., 2016). We also included healthy controls that underwent spinal cord MRI within previous studies conducted at the Queen Square MS Centre, University College London (Brownlee et al., 2017; Kearney et al., 2014); the latter dataset was included in order to characterize physiological spinal cord atrophy rates and consequently compare to pathological rates.

2.3.1. Test dataset

We used previously acquired 3D T1-w images of the spinal cord with 1 mm isotropic voxel (Yiannakas et al., 2016) using a 3 T Philips Achieva MRI system with RF dual-transmit technology (Philips Medical Systems, Best, Netherlands). These data were acquired twice with repositioning of the subjects in-between acquisitions, on 8 healthy controls (mean age 33.5 ± 6.7 years). Afterwards, we performed over the same 8 subjects the spinal cord segmentation using three well-established techniques: a semi-automatic delineation method of the CSA, using JIM 6, and two fully-automatic segmentation methods using PropSeg and DeepSeg algorithms available with the SCT (De Leener, Lévy, et al., 2017; Gros et al., 2019). From these three segmentation techniques, we obtained PCVC with GBSI.

2.3.2. Multiple sclerosis trial data

INFORMS is a phase 3, randomised, double-blind, placebo-controlled clinical trial that included 970 primary progressive MS (PPMS) patients with an EDSS score between 3.5 and 6 recruited from September 2008 to August 2011 from 148 different sites across the world, using 1.5 T and 3 T MRI scanners. The trial compared oral fingolimod to placebo, but failed to demonstrate any efficacy on both clinical (e.g., disability progression) and radiological outcomes (e.g., brain and spinal cord atrophy) (Lublin et al., 2016; Yaldizli et al., 2015).

From the original trial population (n = 970) (Lublin et al., 2016), we included only patients with dedicated T1-w spinal cord scans (1 mm

isotropic isometric voxel) at baseline and 1-year follow-up. The inclusion of patients with spinal cord scans at baseline and 1-year follow-up visits was decided because of the previous evidence that cord atrophy was non-linear in this cohort over the years (Yaldizli et al., 2015), and thus it was preferable to include a homogeneous population, representative of PPMS recruited for a hypothetical phase 2 clinical trial of 1-year duration using spinal cord atrophy as the main outcome measure. Two independent raters (MM and FP) performed a quality check of the images. MM checked pre-processed images (signal-to-noise ratio, spinal cord coverage, image contrast, differences between time-points). FP checked post-processed results, after registration between time-points, straightening and longitudinal bias field correction, and prior to computing GBSI. Each INFORMS site was then independently classified as providing poor, average or good quality data. During the review, 49 subjects (5.3%) were excluded due to sub-optimal image quality (e.g. poor image contrast or motion-related artifacts) and 251 (25.9%) were excluded because no dedicated longitudinal spinal cord acquisitions were available. The overall agreement for this visual quality check between raters was 97.3% (Cohen's kappa = 0.949). Considering that there was 100% agreement between raters when using data from sites providing good quality images, for statistical purposes, we opted to group all sites into 1) sites providing good quality data (28 sites, 131 patients), and 2) and sites providing average-low quality data (102 sites, 479 patients) (**see** Fig. 2).

2.4. Single site healthy control group

We included 52 healthy controls (age = 28.6 years; female sex = 53.5%) with spinal cord scans acquired 1-year apart as a comparison group for spinal cord atrophy measurements. Scans were obtained from previous MRI studies conducted at the UCL Queen Square Institute of Neurology, London, UK (Brownlee et al., 2017; Kearney et al., 2014). All of them underwent a dedicated 3D T1-w 1 mm isotropic voxel acquisition using a 3 T Philips Achieva MRI system with RF dual-transmit technology (Philips Medical Systems, Best, Netherlands). Images were acquired and processed as previously described in this paper for the INFORMS trial. We used this group as reference group for subtracting physiological spinal cord atrophy and, thus, for computing pathological spinal cord atrophy in the INFORMS data. We expected atrophy values close to 0 for both measures (CSA and GBSI).

2.5. Ethics statement

Consent forms were approved by the relevant institutional review boards, and all patients gave written informed consent.

2.6. Statistical methods

To evaluate the reproducibility of GBSI measurements obtained with different segmentation methods, we calculated the intraclass correlation coefficient (ICC) for GBSI obtained from JIM and SCT segmentations using PropSeg (De Leener, Lévy, et al., 2017) and DeepSeg (Gros et al., 2019) (on the test dataset).

To evaluate the agreement between segmentation and registration measurements (on the INFORMS dataset), CSA and GBSI methods were included in a linear regression model. Also, CSA and GBSI were compared using Cohen's *d* effect size, estimating the mean difference between measurements.

To evaluate the spinal cord atrophy rates for segmentation and

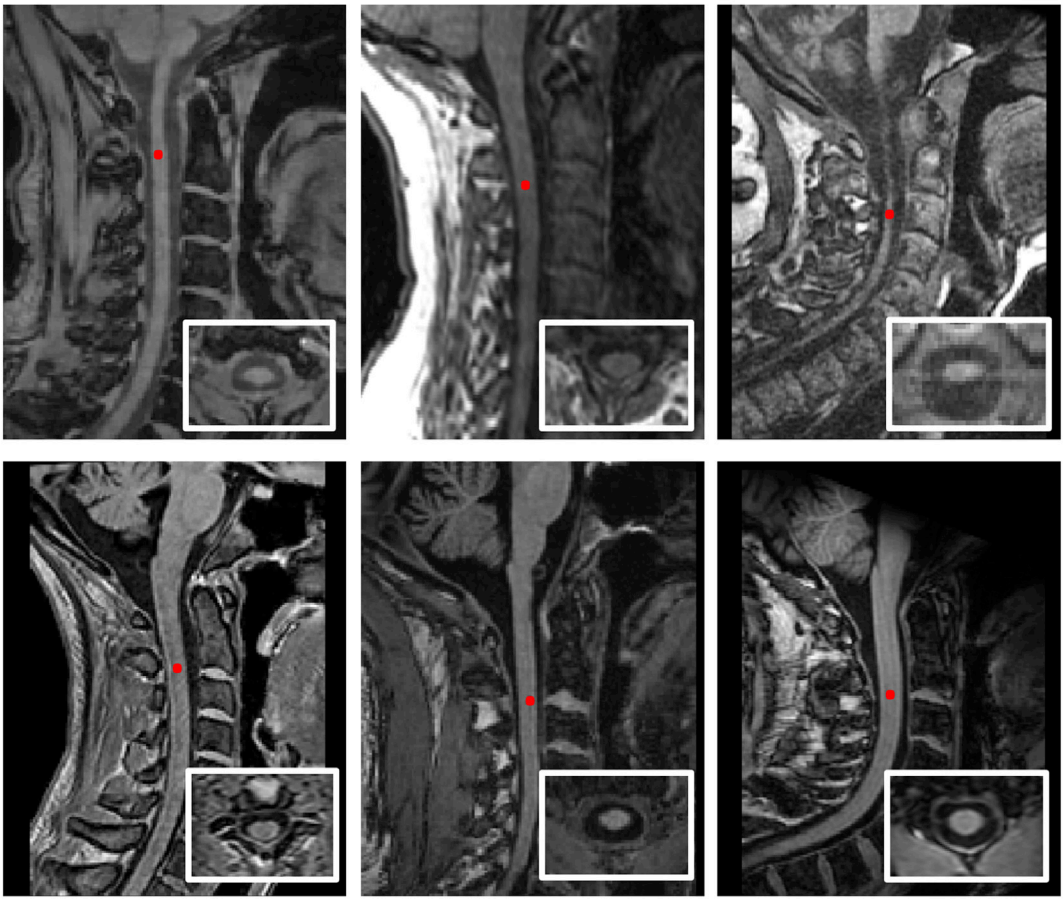

Fig. 2. Example of poor and good quality scans within the INFORMS cohort. First row shows images classified as of poor quality, second row shows example images of good quality.

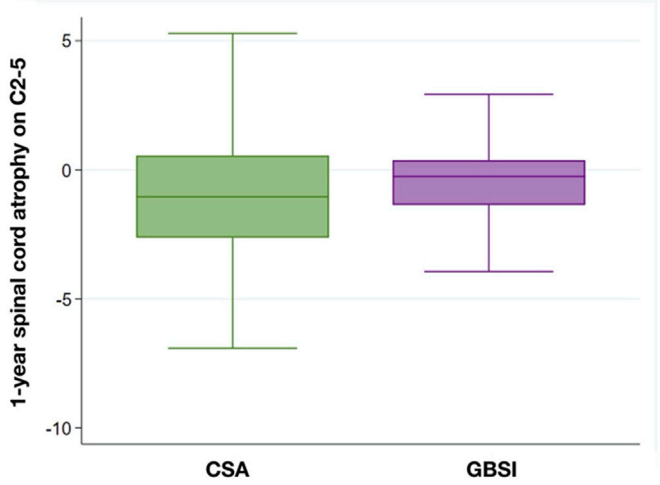

Fig. 3. Box-and-Whisker plots show the 1-year spinal cord atrophy at C2-5 level with CSA (Cross-Sectional Area) and GBSI (Generalised Boundary Shift Integral).

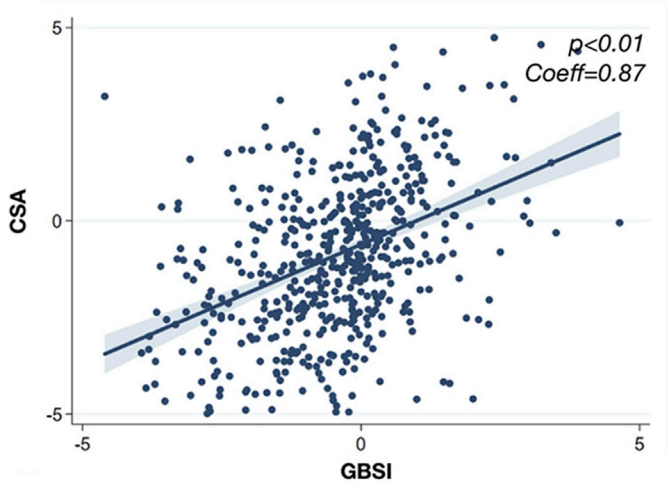

Fig. 4. Scatter plot shows association between 1-year spinal cord atrophy measurements obtained at C2-5 level using CSA (Cross-sectional Area) and GBSI (Generalised Boundary Shift Integral). P-value and coefficient is reported from linear regression model.

registration measurements, mean PCVC and standard deviation were calculated for CSA and GBSI methods on the INFORMS dataset and in controls; mean PCVC and standard deviation were also calculated in INFORMS sites providing good quality data. Differences in PCVC measurements between PPMS and controls (independent variable, using controls as the reference group) were estimated using linear regression models including CSA and GBSI in turn as the dependent variable, and country and site as covariates. Results are presented as coefficients (Coeff) (reflecting the change in PCVC that corresponds to PPMS, when compared with physiological spinal cord loss in controls), 95% confidence intervals (95%CI) and p-values (Schneider et al., 2010).

To evaluate the measurement precision of CSA and GBSI methods (from the INFORMS dataset), we computed the sample size needed to detect treatment effect from spinal cord atrophy in a hypothetical clinical trial of MS patients, using the formula $n = \frac{2(Z_\alpha + Z_{1-\beta})^2\sigma^2}{\Delta^2}$, where n is the required sample size per treatment arm in 1:1 controlled trials, Z_α and $Z_{1-\beta}$ are constants (set at 5% alpha-error and 80% power, respectively), σ is the standard deviation of the measurement (PCVC for CSA or GBSI) and Δ the estimated effect size (Altmann et al., 2009). With a conservative approach, the treatment effect was defined as the variation in PCVC between PPMS and controls, as estimated by Coeff from the linear regression models (Cawley et al., 2018; Moccia and Prados, 2019). Different treatment effects were considered (e.g., 30%, 60% and 90%), which were representative of the variation in spinal cord atrophy measurements expected in MS patients when compared with the physiological loss in spinal cord size. The standard deviation for each group was included in the sample size formula. Calculations were performed first in the whole PPMS cohort and then by selecting the sites which provided good quality images only.

Significance level was set at $p < 0.05$. Statistical analyses were performed using Stata 15.0 (College Station, Texas, US).

All the reported longitudinal atrophy values are in percentage units.

3. Results

3.1. Test dataset

On the test dataset, we observed moderately similar atrophy values for GBSI independently of the segmentation technique used as input (GBSI from JIM segmentation = −0.38 ± 1.48, GBSI from SCT Propseg segmentation = −0.40 ± 3.07, GBSI from SCT Deepseg segmentation = −0.86 ± 3.66, ICC(JIM, SCT Propseg and SCT Deepseg) = 0.73).

3.2. INFORMS dataset and controls

For the INFORMS dataset, the final cohort included 610 PPMS patients (62.8% of the trial population), coming from 130 sites across 18 countries. As a comparison group, we included 52 healthy controls from one site.

In PPMS patients from all sites, we observed spinal cord loss during the course of 1-year (CSA = −1.33 ± 3.62; GBSI = −0.50 ± 1.60) (Fig. 3). Spinal cord atrophy measurements obtained with CSA were correlated to those obtained with GBSI (Coeff = 0.86; 95% CI = 0.71, 1.02; $p < 0.01$) (Fig. 4), with relatively small difference between the two groups (Cohen's d effect size = 0.17, corresponding to 17% difference between measurements). GBSI resulted in lower standard deviation (1.60% vs 3.62%), and lower median absolute deviation than CSA (1.32% vs 2.50%). When including INFORMS good image quality sites only, spinal cord atrophy was more pronounced and associated with lower standard deviation for both methods (CSA = −1.58 ± 2.95; GBSI = −0.79 ± 1.39).

In healthy controls, PCVC remained essentially stable over 1 year (CSA = −0.53 ± 5.24; GBSI = −0.11 ± 2.72). GBSI provided a smaller mean and standard deviation than CSA.

On the linear regression model adjusted by country and site of MRI acquisition, there was no significant difference in 1-year PCVC between PPMS and healthy controls using CSA (all sites: Coeff = −0.80; 95%CI = −1.87, 0.26; p = 0.14; good image quality sites: Coeff = −1.05; 95%CI = −2.23, 0.12; p = 0.08); using GBSI, there was a significantly faster rate of spinal cord atrophy in PPMS, when compared with healthy controls (all sites: Coeff = −0.62; 95%CI = −1.10, −0.13; p = 0.01; good image quality sites: Coeff = −0.84; 95%CI = 1.43, −0.25; $p < 0.01$). Sample size estimates were consistently lower for GBSI, when compared with CSA, especially for good quality sites (Table 1).

4. Discussion

Our results showed that a registration-based measurement of spinal cord atrophy (GBSI) improves spinal cord atrophy measurement precision and sensitivity to change in longitudinal studies. GBSI exceeded the sensitivity established by the present gold standard method for measuring longitudinal spinal cord atrophy (i.e., segmentation methods). The better performance using GBSI is directly related to the greater measurement precision (as indirectly shown by the lower standard deviation). This is because GBSI is able to derive PCVC values directly from small intensity changes between images at the cord boundaries,

Table 1
Sample size estimates for CSA (Cross-sectional Area) and GBSI (Generalised Boundary Shift Integral). Sample size estimates are reported for CSA and GBSI. Number of included patients for the analysis, and coefficient of spinal cord atrophy with standard deviation used in the sample size formula are also reported. Different effect size has been hypothesized (30%, 60% and 90%).

	CSA		GBSI	
Effect size	All sites *N=479* *Coeff=-0.80* *SD=* 3.62	Good image quality sites *N=131* *Coeff=-1.05* *SD=* 2.95	All sites *N=479* *Coeff=-0.62* *SD=* 1.60	Good image quality sites *N=131* *Coeff=-0.84* *SD=* 1.39
30%	3567	1375	1160	408
60%	892	344	290	102
90%	396	153	129	45

accounting for partial volume effects in these regions which are critical for measuring changes. Moreover, the use of the probabilistic XOR (pXOR) region for weighting the boundary shift integral benefits the calculation over specific tissue boundaries, excluding voxels in areas that might reduce its sensitivity (e.g., voxels with partial volume with CSF or the centre voxels of the spinal cord). This is particularly relevant in the spinal cord where there are extremely close-fitting surfaces that can be affected by changes even in a small number of voxels.

Conventional segmentation-based methods (e.g, CSA) rely on the difference between the mean of the areas obtained from the hard segmentation at each time-point. This approach is not considering partial volume averaging effects and can introduce greater variability, especially in acquisitions with large voxel sizes or between scans with different intensity scales. This fluctuation in CSA differences is also a potential explanation for GBSI's superior performance, systematically having smaller standard deviations and sample size estimates than CSA.

The correlation between CSA and GBSI was very high (see Fig. 4, Coeff = 0.87 and $p < 0.01$), showing that they had an excellent agreement, with small mean difference. Hence, we can consider that GBSI is measuring similar cord change (or atrophy) as with CSA, but with smaller standard deviation (and higher sensitivity).

Test results using three segmentation techniques showed that GBSI is robust and able to measure similar levels of atrophy, independently of the segmentation method used (JIM, SCT Propseg or SCT DeepSeg). This is a consequence of computing atrophy as the intensity difference of the spinal cord boundary. The standard deviation differences come from the changes in the outer and inner border of the pXOR mask. The manual delineation using ASM from JIM tends to generate a slightly thinner pXOR mask than PropSeg from SCT (Fig. 5). However, test dataset ICC values for GBSI were lower than previously published for CSA (Yiannakas et al., 2016). Overall, we demonstrated the feasibility to obtain longitudinal spinal cord atrophy rates with GBSI using JIM, SCT PropSeg or SCT DeepSeg. Future work could consider analysing the stability of the method using different spinal cord segments, a wider range of spinal cord segmentation methods and different fields-of-view.

Future longitudinal observational studies and clinical trials in MS can benefit from GBSI for spinal cord atrophy calculation, as an important surrogate marker of disability progression (Ciccarelli et al., 2019; Moccia and Ruggieri, 2019). Our sample size estimates for 1-year spinal cord atrophy with GBSI are of the same magnitude as those required for brain atrophy, the current gold standard primary outcome measure in phase 2 clinical trials for progressive MS (Chataway et al., 2014; Fox et al., 2018; Cambron et al., 2019). For instance, the sample size to detect a 60% treatment effect on spinal cord atrophy in 1-year is 290 subjects for GBSI and three-fold larger with CSA. Further improvements could be achieved by including only sites providing good quality images (e.g., standard acquisition protocol, centralized MRI acquisition).

Unfortunately, clinical variables were not available and, thus, it was not possible to assess the potential effect size of the intervention. In the future, it would be interesting to assess also the sensitivity to change in the presence of treatment. Change in MS lesion intensities might have an impact on GBSI estimates, and future work could address the impact of MS lesions and how to reduce it. Another limitation was the use a control group from a single centre/scanner; to the best of our knowledge, there is no existing dataset of healthy controls with sample size and MRI acquisitions fully comparable to INFORMS. In the absence of that, we used our single centre/scanner healthy control group to obtain more conservative atrophy estimates (and subsequent sample size calculation) than would have been the case when assuming zero atrophy with no standard deviation, which would have led to overly optimistic sample size estimates. Finally, this control population is over ten-fold smaller than MS patients, possibly explaining the larger standard deviation.

The acquired voxel size is the main limiting factor that determines the degree of precision of GBSI, where small and isotropic voxels are the most suitable to use. Anisotropic voxel sizes (e.g., with 5 mm or more slice thickness, as frequently implemented in spinal cord protocols) will introduce inconsistencies when using the present pipeline. As this is a longitudinal approach, it is quite plausible that if each acquisition is performed with a different positioning of the slices compared to the actual spinal cord segments, even simply because of curvature, slices between time-points could hardly be matched, consequently impacting on the registration step and the final shift integral calculation. Thus, GBSI requires a standard high-quality isotropic T1-w image protocol for spinal cord MRI. Nowadays this is achievable thanks to a consensus multi-vendor dedicated protocol that has been developed between 30 MRI international centres and made publicly available at the website http://www.spinalcordmri.org (Protocols' section). This protocol eases the adoption of the most standard spinal cord MRI acquisitions and, consequently, the use of GBSI. Finally, GBSI has been developed so far for T1-w images, which limits its applicability to the spinal cord, where this type of acquisition is frequently not available. The GBSI pipeline includes several intensity corrections (bias field correction, longitudinal symmetric bias field correction, intensity normalization and use of a clipping window) to obtain a direct estimate of tissue change between timepoints; despite these efforts to equalise the images, we still might find some residual noise due to the underlying signal. As future work, we aim to adapt GBSI to support other types of spinal cord images.

5. Conclusion

In this work, we introduced a new pipeline based on the latest iteration of BSI for computing longitudinal atrophy in the spinal cord and

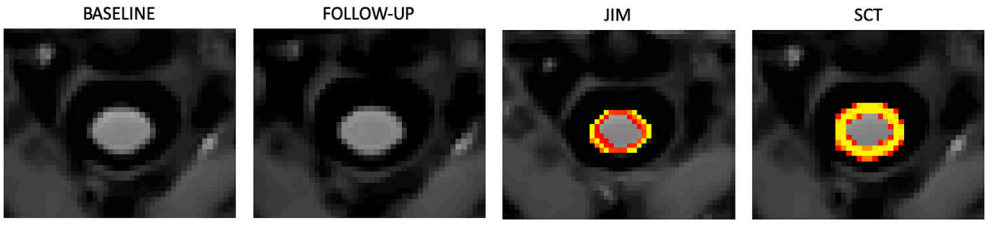

Fig. 5. Example of pXOR masks obtained from baseline and follow-up segmentations using Active Surface Modelling from JIM, or PropSeg from SCT (Spinal Cord Toolbox).

compared its results with the commonly used segmentation-based spinal cord atrophy measurement (numerical difference of mean CSA between time-points). We demonstrated that GBSI, a registration-based technique, is a sensitive, quantitative and objective measure of longitudinal tissue volume changes in the spinal cord. The GBSI pipeline presented in this work is repeatable and reproducible and could become the preferred method for computing longitudinal spinal cord atrophy in clinical trials and observational studies.

Authors contributions

Ferran Prados: Conceptualization, Methodology, Software, Investigation, Data curation, Validation, Writing- Original draft preparation, Writing - review & editing, Funding Acquisition. **Marcello Moccia**: Conceptualization, Methodology, Investigation, Data curation, Visualization, Writing- Original draft preparation, Writing - review & editing. **Aubrey Johnson**: Data curation, Visualization, Validation, Investigation. **Marios Yiannakas:** Methodology, Writing- Original draft preparation. **Francesco Grussu:** Methodology, Writing- Original draft preparation, Writing - review & editing. **Manuel Jorge Cardoso:** Methodology, Writing- Original draft preparation. **Olga Ciccarelli:** Methodology, Resources, Supervision, Writing- Original draft preparation. **Sebastien Ourselin:** Conceptualization, Supervision. **Frederik Barkhof**: Resources, Methodology, Supervision, Writing- Original draft preparation, Writing - review & editing. **Claudia Wheeler-Kingshott:** Conceptualization, Methodology, Supervision, Writing- Original draft preparation, Writing - review & editing.

Acknowledgements

FP has a Non-Clinical Postdoctoral Guarantors of Brain fellowship. This research was funded by the UK MS Society (programme grant number 984). This study was supported by researchers at the National Institute for Health Research University College London Hospitals Biomedical Research Centre (FB and OC). This project has received funding under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 634541 and from the Engineering and Physical Sciences Research Council (EPSRC) EP/R006032/1, funding FG.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.116489.

References

Altmann, D.R., et al., 2009. Sample sizes for brain atrophy outcomes in trials for secondary progressive multiple sclerosis. Neurology 72 (7), 595–601.
Antonescu, F., et al., 2018. A review of cervical spine MRI in ALS patients. J. Med. Life 11 (2), 123–127.
Brownlee, W.J., et al., 2017. Association of asymptomatic spinal cord lesions and atrophy with disability 5 years after a clinically isolated syndrome. Mult. Scler. 23 (5), 665–674.
Buades, A., Coll, B., Morel, J.-M., 2005. A Non-local Algorithm for Image Denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). https://doi.org/10.1109/cvpr.2005.38. Available at:
Cambron, M., et al., 2019. Fluoxetine in Progressive Multiple Sclerosis: the FLUOX-PMS Trial. Multiple sclerosis, 1352458519843051.
Cawley, N., et al., 2018. Spinal cord atrophy as a primary outcome measure in phase II trials of progressive multiple sclerosis. Multiple Sclerosis 24, 932–941.
Chataway, J., et al., 2014. Effect of high-dose simvastatin on brain atrophy and disability in secondary progressive multiple sclerosis (MS-STAT): a randomised, placebo-controlled, phase 2 trial. The Lancet 383 (9936), 2213–2221.
Ciccarelli, O., et al., 2019. Spinal cord involvement in multiple sclerosis and neuromyelitis optica spectrum disorders. Lancet Neurol. 18 (2), 185–197.
De Leener, B., Lévy, S., et al., 2017. SCT: spinal Cord Toolbox, an open-source software for processing spinal cord MRI data. Neuroimage 145 (Pt A), 24–43.
De Leener, B., Mangeat, G., et al., 2017. Topologically preserving straightening of spinal cord MRI. J. Magn. Reson. Imaging: JMRI 46 (4), 1209–1219.
Denecke, C.K., Aljović, A., Bareyre, F.M., 2019. Combining molecular intervention with in vivo imaging to untangle mechanisms of axon pathology and outgrowth following spinal cord injury. Exp. Neurol. 318, 1–11.
El Mendili, M.M., et al., 2019. Spinal cord imaging in amyotrophic lateral sclerosis: historical concepts-novel techniques. Front. Neurol. 10, 350.
Fox, R.J., et al., 2018. Phase 2 trial of ibudilast in progressive multiple sclerosis. N. Engl. J. Med. 379 (9), 846–855.
Freeborough, P.A., Fox, N.C., 1997. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. IEEE Trans. Med. Imaging 16 (5), 623–629.
Gros, C., et al., 2019. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. Neuroimage 184, 901–915.
Horsfield, M.A., et al., 2010. Rapid semi-automatic segmentation of the spinal cord from magnetic resonance images: application in multiple sclerosis. Neuroimage 50 (2), 446–455.
Jones, D.K., Basser, P.J., 2004. "Squashing peanuts and smashing pumpkins": how noise distorts diffusion-weighted MR data. Magn. Reson. Med. 52 (5), 979–993.
Kalkers, N.F., et al., 2002. The effect of the neuroprotective agent riluzole on MRI parameters in primary progressive multiple sclerosis: a pilot study. Mult. Scler. 8 (6), 532–533.
Kapoor, R., et al., 2010. Lamotrigine for neuroprotection in secondary progressive multiple sclerosis: a randomised, double-blind, placebo-controlled, parallel-group trial. Lancet Neurol. 9 (7), 681–688.
Kappos, L., et al., 2018. Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. The Lancet 391 (10127), 1263–1273.
Kearney, H., et al., 2014. Improved MRI quantification of spinal cord atrophy in multiple sclerosis. J. Magn. Reson. Imaging: JMRI 39 (3), 617–623.
Leary, S.M., et al., 2003. Interferon beta-1a in primary progressive MS: an exploratory, randomized, controlled trial. Neurology 60 (1), 44–51.
Leung, K.K., et al., 2012. Consistent multi-time-point brain atrophy estimation from the boundary shift integral. Neuroimage 59 (4), 3995–4005.
Leung, K.K., et al., 2010. Robust atrophy rate measurement in Alzheimer's disease using multi-site serial MRI: tissue-specific intensity normalization and parameter selection. Neuroimage 50 (2), 516–523.
Lewis, E.B., Fox, N.C., 2004. Correction of differential intensity inhomogeneity in longitudinal MR images. Neuroimage 23 (1), 75–83.
Lin, X., et al., 2003. Spinal cord atrophy and disability in multiple sclerosis over four years: application of a reproducible automated technique in monitoring disease progression in a cohort of the interferon beta-1a (Rebif) treatment trial. J. Neurol. Neurosurg. Psychiatry 74 (8), 1090–1094.
Lublin, F., et al., 2016. Oral fingolimod in primary progressive multiple sclerosis (INFORMS): a phase 3, randomised, double-blind, placebo-controlled trial. The Lancet 387 (10023), 1075–1084.
Moccia, M., Ruggieri, S., et al., 2019. Advances in spinal cord imaging in multiple sclerosis. Ther. Adv. Neurol. Disord. 12, 1756286419840593.
Moccia, M., de Stefano, N., Barkhof, F., 2017. Imaging outcomes measures for progressive multiple sclerosis trials. Multiple Sclerosis 23, 1614–1626.
Moccia, M., Prados, F., et al., 2019. Longitudinal spinal cord atrophy in multiple sclerosis using the generalized boundary shift integral. Ann. Neurol. https://doi.org/10.1002/ana.25571. Available at:
Modat, M., et al., 2014. Global image registration using a symmetric block-matching approach. J. Med. Imaging 1 (2), 024003.
Montalban, X., et al., 2017. Ocrelizumab versus placebo in primary progressive multiple sclerosis. N. Engl. J. Med. 376 (3), 209–220.
Montalban, X., et al., 2009. Primary progressive multiple sclerosis diagnostic criteria: a reappraisal. Mult. Scler. 15 (12), 1459–1465.
Prados, F., et al., 2016. Fully automated grey and white matter spinal cord segmentation. Sci. Rep. 6, 36151.
Prados, F., et al., 2015. Measuring brain atrophy with a generalized formulation of the boundary shift integral. Neurobiol. Aging 36 (Suppl. 1), S81–S90.
Prados, F., Barkhof, F., 2018. Spinal cord atrophy rates. Ready for prime time in multiple sclerosis clinical trials? Neurology 91, 157–158.
Reuter, M., et al., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61 (4), 1402–1418.
Schneider, A., Hommel, G., Blettner, M., 2010. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch. Arzteblatt 107 (44), 776–782.
Schott, J.M., et al., 2010. Reduced sample sizes for atrophy outcomes in Alzheimer's disease trials: baseline adjustment. Neurobiol. Aging 31 (8), 1452–1462, 1462.e1–2.
Smith, S., et al., 2000. Siena — normalised accurate measurement of longitudinal brain change. Neuroimage 11 (5), S659. https://doi.org/10.1016/s1053-8119(00)91589-1. Available at:
Smith, S., et al., 2001. SIENA: single and multiple time point brain atrophy analysis. Neuroimage 13 (6), 250. https://doi.org/10.1016/s1053-8119(01)91593-9. Available at:
Stroman, P.W., et al., 2014. The Current State-Of-The-Art of Spinal Cord Imaging: Methods. Neuroimage vol. 84, 1070–1081.
Tur, C., et al., 2018. Assessing treatment outcomes in multiple sclerosis trials and in the clinical setting. Nat. Rev. Neurol. 14 (2), 75–93. https://doi.org/10.1038/nrneurol.2017.171. Available at:
Tristán-Vega, A., Aja-Fernández, S., Westin, C.-F., 2012. Least squares for diffusion tensor estimation revisited: propagation of uncertainty with Rician and non-Rician signals. Neuroimage 59 (4), 4032–4043. https://doi.org/10.1016/j.neuroimage.2011.09.074. Available at:

Tustison, N.J., et al., 2010. N4ITK: improved N3 bias correction with robust B-spline approximation. 2010 IEEE international symposium on biomedical imaging: from nano to macro. Available at: https://doi.org/10.1109/isbi.2010.5490078.

Wheeler-Kingshott, C.A., et al., 2014. The current state-of-the-art of spinal cord imaging: applications. Neuroimage 84, 1082–1093.

Yaldizli, Ö., et al., 2015. Brain and cervical spinal cord atrophy in primary progressive multiple sclerosis: results from a placebo-controlled phase III trial (INFORMS). Available at: https://onlinelibrary.ectrims-congress.eu/ectrims/2015/31st/116684/oezguer.yaldizli.brain.and.cervical.spinal.cord.atrophy.in.primary.progressive.html.

Yiannakas, M.C., et al., 2012. Feasibility of grey matter and white matter segmentation of the upper cervical cord in vivo: a pilot study with application to magnetisation transfer measurements. Neuroimage 63 (3), 1054–1059.

Yiannakas, M.C., et al., 2016. Fully automated segmentation of the cervical cord from T1-weighted MRI using PropSeg: application to multiple sclerosis. NeuroImage: Clinical 10, 71–77.