Expert Variability and Deep Learning Performance in Spinal Cord Lesion Segmentation for Multiple Sclerosis Patients

Ricky Walsh, Cédric Meurée, Anne Kerbrat, Arthur Masson, Burhan Rashid Hussein, Malo Gaubert, Francesca Galassi, Benoit Combés

Expert Variability and Deep Learning Performance in Spinal Cord Lesion Segmentation for Multiple Sclerosis Patients

Ricky Walsh[1,*,†], Cédric Meurée[1,*,†], Anne Kerbrat[1,2], Arthur Masson[1]
Burhan Rashid Hussein[1], Malo Gaubert[1,3], Francesca Galassi[1,**], Benoit Combès[1,**]
[1] *Univ Rennes, Inria, CNRS, Inserm, IRISA UMR 6074, Empenn ERL U1228, Rennes, France*
[2] *Rennes University Hospital (CHU), Department of Neurology, Rennes, France*
[3] *Rennes University Hospital (CHU), Department of Neuroradiology, Rennes, France*
[*,**] Authors contributed equally
[†] Corresponding authors: ricky.walsh@inria.fr; cedric.meuree@inria.fr

***Abstract*—Multiple sclerosis (MS) patients often present with lesions in spinal cord magnetic resonance (MR) volumes. However, accurately detecting these lesions is challenging and prone to inter- and intra-rater variability. Deep learning-based methods have the potential to aid clinicians in detecting and segmenting MS lesions, but can also be affected by rater variability.**

This study assesses the inter- and intra-rater variability in manual segmentation of spinal cord lesions, and evaluates raters and a state-of-the-art nnU-Net model against a ground truth (GT) segmentation of a senior expert. Four experts segmented twelve spinal cord MR volumes from six patients twice, at a time distance of two weeks.

Considerable inter- and intra-rater variability were observed, with the total number of detected lesions ranging from 28 to 60, depending on the rater. Moreover, the segmented volumes of individual lesions varied substantially between raters. All raters and the model achieved high precision when evaluated against the senior expert GT, but sensitivity was notably lower. These results motivate the need for more sensitive automated methods to aid clinicians in lesion detection, and suggest that consideration should be given to inter-rater variability when training and evaluating automated methods.

***Index Terms*—Multiple sclerosis, spinal cord, magnetic resonance imaging, lesion segmentation, inter-rater variability, intra-rater variability, deep learning, automated segmentation**

I. Introduction

Multiple Sclerosis (MS) is a chronic inflammatory demyelinating disease of the central nervous system [1]. Magnetic Resonance (MR) imaging of the brain and spinal cord is used to identify MS lesions, providing an essential aid to clinicians in the diagnosis and prognosis of this disease [2]. In particular, MS lesions in the spinal cord are prognostic of short-term disability progression [3] and thus can indicate the need for more aggressive treatments.

The identification of MS lesions on a given MR image is a complex and mentally demanding task and can lead to an underestimation of disease activity, even for experienced radiologists and neurologists. The quality of a segmentation is influenced, among other factors, by the characteristics of the lesions, the experience of the clinician examining the image, the amount of information available (e.g., the number of scans or clinical information), and image quality. The quality of MR images of the spinal cord is particularly affected by artifacts from patient movement or respiratory and cardiac cycles [4], which can introduce uncertainty as to whether certain hyper-intense regions in the image are truly lesions or merely artifacts. Even after determining that a lesion exists, delineating the lesion boundaries is also complicated by these artifacts and partial volume effects in the image, and is particularly difficult for diffuse lesions even in the absence of artifacts. Inter- and intra-rater variability in assessing disease involvement in spinal cord MR volumes can arise from these challenges and could have important clinical consequences if it affects the treatment selected for the patient.

Automated methods have been proposed to aid radiologists and neurologists in the challenging task of identifying lesions in the spinal cord and to reduce inter- and intra-rater variability [5]. Methods based on deep learning are now the state of the art in this regard. However, these methods rely on learning lesion patterns from a set of images where lesions have been delineated by experts. Inter-rater variability can therefore affect these models in two ways. Firstly, the models may learn to recreate the segmentation patterns of the individual raters rather than the "true" segmentation, leading to sub-optimal results [6]. Secondly, to evaluate the model, its predictions are compared to some expert "ground truth" annotation, and variability in these annotations could result in an underestimate or overestimate of model performance. It is therefore important to assess the extent of rater variability, both to explore the potential added value of automated methods and

This study was partly funded by the French doctoral program in artificial intelligence (reference: ANR-20-THIA-0018). The study was supported by the French National Research Agency (ANR) within the France 2030 program, 3rd PIA (reference ANR-21-RHUS-0014). Data collection was supported by a grant provided by the French State and handled by ANR within the France 2030 program (reference ANR-10-COHO-002 OFSEP).

to understand the potential consequences for the training and evaluation of deep learning-based methods.

The main objectives of this study are to:

- assess inter- and intra-rater variability in segmentations of MS lesions in spinal cord MR volumes, and
- evaluate the relative performance of experts and assess a state-of-the-art automated method in this context.

Several studies have examined inter-rater variability on annotations of MS lesions in brain MR volumes [7]–[9], and Gros *et al.* [5] reported the variability in Dice scores of several raters compared to a majority consensus in spinal cord MR volumes. The current study builds upon these previous works to explore and quantify the variability with regard to spinal cord MR imaging. In particular, we include an intra-rater analysis, an evaluation of raters with respect to a ground truth adjudicated by a senior expert, and a deeper analysis of variability with respect to lesion-wise sensitivity and precision. Furthermore, we evaluate an automated segmentation method based on the nnU-Net [10] framework within the context of this rater variability. Finally, we compare automatically generated consensuses to an adjudicated segmentation of a senior expert. In the absence of an expert consensus or adjudication, studies on inter-rater variability often use automated consensus methods, e.g., Simultaneous Truth and Performance Level Estimation (STAPLE) [11] or taking a majority vote on each voxel. We examine how well these consensus methods approximate the segmentation of a senior expert in the context of spinal cord MS lesions.

II. Experimental Setup

A. Dataset

The data for this study were gathered from the OFSEP (Observatoire Français de la Sclérose en Plaques - French MS registry) database [12]. The data consisted of MR volumes of six patients diagnosed with MS. Each patient had one sagittal T2-weighted (T2-w) volume for the upper section of the spinal cord (12-13 vertebrae), and another for the lower spinal cord. Furthermore, three of the six patients also had corresponding Short inversion Time Inversion Recovery (STIR) volumes available from the same acquisition session [13]. Three subjects were imaged on a Siemens Aera (1.5T) scanner, and the scans of the other subjects came from a Siemens Avanto (1.5T), a Siemens Spectra (3T) and a General Electric Optima MR450w (1.5T).

The data were selected based on an expert visual inspection with the following criteria: the MR volumes should not present strong acquisition artifacts (e.g., noise, motion), and the selected data should be representative of a range of potentially encountered clinical cases, i.e., ranging from a patient with no visible lesions to patients with a high lesion load. Based on the segmentation of the senior expert described in Sec. II-D1, three of the twelve MR volumes contained no lesions (i.e., both volumes of one patient and one volume of another). In the remaining nine volumes, a total of 63 lesions were segmented by the senior expert, with a mean volume of 599.9mm^3. Each of these images contained between 2 and 15 lesions, with a minimum and maximum volume of 0.4mm^3 and 3056.6mm^3, respectively. The voxel sizes of the T2-w acquisitions were (3.04 $\pm$ 0.56, 0.53 $\pm$ 0.14, 0.53 $\pm$ 0.14) mm^3, and (2.90 $\pm$ 0.15, 0.61 $\pm$ 0.18, 0.61 $\pm$ 0.18) mm^3 for the STIR volumes, reported as *mean* $\pm$ *standard deviation*. The MR volumes were acquired on four different scanners from two vendors.

B. Expert Annotations

Four experts were asked to identify and delineate the lesions on each of the 12 sagittal T2-w volumes, with the aid of STIR volumes when available ($3/6$ patients). The raters comprised three radiologists and one neurologist, and each had approximately five years of experience. The segmentations were performed using the ITK-SNAP software [14], all raters having been previously trained to use it for this particular task. The raters were asked to repeat this segmentation task on the same MR volumes in a second session two weeks later, resulting in eight rater annotations per MR volume.

C. Automated Segmentation

A segmentation model was trained using the nnU-Net framework [10] on a separate dataset of 62 sagittal T2-w volumes for training and 13 volumes as a validation set (voxel sizes=(2.81 $\pm$ 0.20, 0.58 $\pm$ 0.12, 0.58 $\pm$ 0.12) mm^3, 44 upper and 18 lower spinal cord volumes, 10 scanner models, 4 brands). The training set contained other volumes from the same four MR scanners on which the six subjects of the current study were imaged. The ground truth (GT) segmentation for each MR volume in the training set was created by one of six experts, including the four experts in the current study. The masks for each MR volume were validated by the senior expert in the current study.

nnU-Net is a self-configuring method based on the U-Net architecture that determines an optimal network topology and hyper-parameters using heuristics and empirical tests during five-fold cross-validation on the training dataset. It has achieved state-of-the-art performance across a variety of medical image segmentation tasks [15].

The 3D volumes were pre-processed by zeroing voxels that do not belong to the spinal cord and applying a re-orientation (RPI, i.e. Right-to-left, Posterior-to-anterior, Inferior-to-superior) and a resampling to voxel sizes of (0.5, 0.5, 0.5) mm^3. Spinal cord masks were obtained using the Spinal Cord Toolbox, as the union of the output masks of the *sct_deepseg_sc* function, dilated with a 5×5×5 structuring element, and the *sct_create_mask* function, which selects 20mm×20mm axial sections around the spinal cord centreline [16]. The union of these two masks was used to avoid occasional errors observed in the individual methods.

The size of input patches to the model was chosen by the nnU-Net framework to be 40×128×496 voxels. Training batches comprised two patches from a single MR volume, where at least one of the patches contained lesion voxels, except for MR volumes with no lesions. The U-Net model chosen by nnU-Net consisted of 5 convolution levels, with

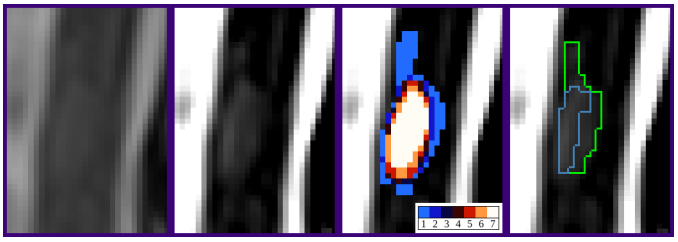

Fig. 1. Example of a lesion in the dataset (at vertebra T6). The bright vertical bands are cerebrospinal fluid surrounding the spinal cord. **Left**: subsection of sagittal T2-w volume. **Centre Left**: contrast adjusted to improve lesion visibility. **Centre Right**: voxel-wise agreement between the eight annotations (four raters with two sessions each). One of the raters did not detect this lesion in one of the sessions. **Right**: predicted segmentation of nnU-Net model (blue) and adjudicated segmentation by senior expert (green).

3×3×3 convolution filters and convolutions with a stride of 2 were used for downsampling. Deep supervision was used during training. The training loss function was the sum of the soft Dice loss and the cross entropy loss. The model was trained for 1000 epochs, with an initial learning rate (lr) equal to 0.01, updated at each epoch by means of the following function: $lr_{epoch+1} = lr_{epoch} * (1 - epoch/1000)^{0.9}$. Output probability maps (p-maps) were thresholded at the default 0.5 to obtain segmentation masks. The resulting model was applied to segment the 12 T2-w volumes in the current study and the segmentation masks were evaluated against a GT segmentation created by a senior expert.

D. Ground Truth

1) Senior Expert Adjudication: Following the rater annotations and model predictions, a total of nine annotations were available for each volume (two annotations per rater and one model prediction). These nine annotations were merged into a single mask of "plausible lesion" voxels for each volume, i.e., any voxel labelled as a lesion voxel in any of the nine annotations was kept as a lesion voxel. A senior neurologist with 10 years of experience was asked to adjudicate on this mask, keeping valid lesion voxels and removing non-lesion voxels. This adjudicated segmentation was used as GT for quantitative evaluations.

Fig. 1 shows an example of a segmented lesion in the dataset, which was detected in seven of the eight rater annotations. The segmentation mask created by the senior expert during the adjudication phase is slightly smaller than accepting all voxels labelled by any of the raters.

2) Automated Consensus: We assessed how well several consensus generation techniques approximated the senior expert adjudication. These included accepting all lesion voxels marked in any of the eight annotations or by the model (*Any Rater*), accepting a lesion voxel if it was segmented by at least five of the nine segmentations per volume (*Majority*), and applying the STAPLE algorithm [11] to the nine segmentations for each volume. STAPLE iteratively estimates the GT by weighting each annotation by its estimated sensitivity and precision, and it is one of the most popular automated consensus generation methods [17]. One of the assumptions of STAPLE is that the annotations are independent of each other, which is not the case in our study as we have two annotations per rater. However, as the number of annotations per rater are equal across raters, it is less serious than if one senior rater were outnumbered by junior raters with correlated annotations, the example given in the original paper on STAPLE [11].

We evaluated these consensus segmentation masks against the senior expert segmentation in the same way as when evaluating the raters and model, which will be described in Sec. II-F. We also assessed the impact on reported Dice similarity scores of the raters and model when using these automated consensuses as GT.

E. Variability Evaluation

The inter- and intra-rater variability was assessed by summarising the number of lesions and lesion volumes segmented by each rater in each session. The Cohen kappa value was computed to evaluate the intra-rater agreement in lesion count between the two sessions for each rater, while the Fleiss kappa calculation was used to assess the inter-rater agreement in each session.

F. Performance Evaluation

Evaluation metrics were based on those presented in Gros *et al.* [5] and the MICCAI16 brain MS segmentation challenge [18]. Voxel-wise sensitivity, precision and Dice similarity coefficient (F1-score) were used to measure the overlap between the senior expert GT and the manual or automatic segmentations. The detection and count of lesions is vital in monitoring MS patients undergoing disease-modifying treatments. In this context, lesion-wise metrics have more value for clinicians than voxel-wise metrics. Therefore, sensitivity, precision and F1-score were also defined at the lesion level.

A GT lesion was treated as correctly detected for the purposes of sensitivity if more than 10% of its voxels were labelled as lesion voxels by a manual or automatic segmentation, i.e., if the voxel-wise sensitivity for this GT lesion exceeded 10%. The 10% threshold follows that used in [18]. Lesion-wise sensitivity is then the ratio of the number of correctly detected GT lesions to the total number of GT lesions.

A predicted lesion was considered to be precise if over 30% of its voxels are truly lesion voxels in the GT, i.e., if the voxel-wise precision of this predicted lesion exceeded 30%. Again, this threshold value follows [18]. To avoid a segmentation being considered more precise if it splits one GT lesion into many predicted lesions, we counted precise predicted lesions only once per GT lesion when calculating precision. Lesion-wise precision is then the ratio of precise predicted lesions to the total number of predicted lesions.

Defining lesion-wise sensitivity and precision separately in terms of only voxel-wise sensitivity and precision, respectively, allows a more straightforward interpretation of the results, i.e., how sensitive and how precise the segmentations were at the lesion level. We further used these lesion-wise sensitivity and precision values in the calculation of lesion-wise F1-score.

III. Results

A. Inter- and Intra-Rater Variability

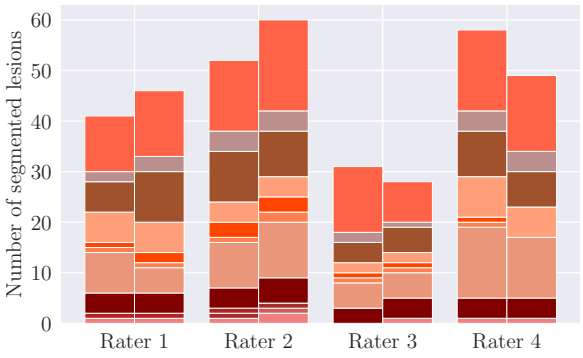

Fig. 2. Number of lesions segmented by each rater in each session. The blocks in each column correspond to individual MR volumes.

1) Lesion Counts: Inter- and intra-rater variability can first be observed in the number of lesions segmented on each MR volume. Differences between raters, as well as between sessions for the same rater, can be observed in the lesion counts in Fig. 2. Concerning intra-rater variability, the number of lesions varied between sessions for all raters. For example, rater 1 segmented six lesions in the upper spinal cord of subject 5 in session 1 but segmented ten lesions in the same MR volume in session 2 (brown blocks in Fig. 2). Comparing the number of lesions in each MR volume across the two sessions for each rater, Cohen kappa values of 0.61, 0.24, 0.50 and 0.50 were obtained for raters 1 to 4, respectively, indicating a high variability from poor to substantial agreement between both sessions, depending on the rater. Inter-rater variability is also apparent, as the total number of lesions ranges from 28 to 60 between raters 3 and 2, respectively. Fleiss kappa values of 0.26 and 0.15 were obtained for sessions 1 and 2, respectively, indicative of the low agreement between raters.

2) Lesion Volumes: Inter- and intra-rater variability is also noticeable in the volume of segmented lesions. Fig. 3 highlights this aspect by presenting the distribution of volumes of individual lesions. Table I complements this figure with summary statistics of the segmented lesion volumes by rater. Raters 2 and 4 tended to delineate smaller regions, with median lesion volumes of 52.6mm^3 and 64.5mm^3, respectively. Raters 1 and 3, on the other hand, had a median lesion volume over double that of rater 2, with values of 113.3mm^3 and 110.3mm^3, respectively.

While Fig. 3 and Table I provide a quantitative evaluation of segmented lesion volumes from the whole dataset (12 MR volumes), a finer analysis at the image level provides nuances to these observations. Indeed, comparisons between raters do not always hold when comparing across MR volumes (Fig. 4). For example, rater 4 segments the highest total lesion volume for subject 3 but the lowest volume for subject 5. Fig. 4 also shows that a higher number of segmented lesions does not necessarily correspond to a higher segmented volume. For instance, rater 2 delineates ten lesions in the upper spinal cord of subject 5 (image 5-U), compared to the six lesions of rater 1. However, the total volume segmented by rater 1 is over three times that of rater 2, indicating a higher disease involvement. This demonstrates that although the number of lesions is a commonly used clinical indicator, it is important to also take the lesion volume into account.

3) Lesion-level Agreement: Fig. 5 shows the level of agreement between the raters' segmentations on individual lesions that appear in the GT adjudicated by the senior expert. 13

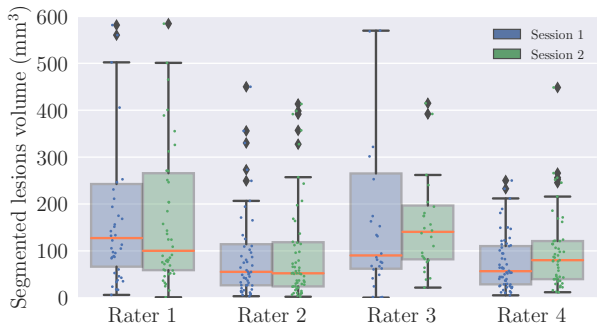

Fig. 3. Lesion volumes segmented by each rater in each session. Each point represents one lesion. This plot is cropped to 600 mm^3, preventing some outliers to appear.

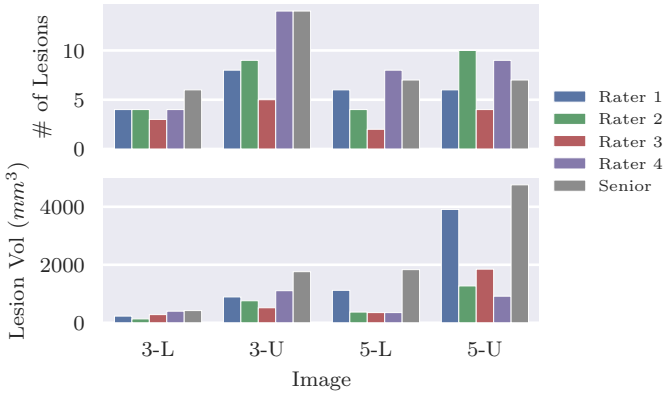

Fig. 4. Number of lesions and total lesion volume segmented per rater for two selected subjects, i.e., four MR volumes. Images 3-L and 3-U correspond to the lower and upper spinal cord, respectively, of subject 3.

TABLE I
SUMMARY STATISTICS OF VOLUMES (MM^3) OF LESIONS DELINEATED BY EACH RATER ACROSS BOTH ANNOTATION SESSIONS.

Rater	Min	Max	Mean	Std. Dev.	Median
Rater 1	1.1	2548.5	246.6	397.3	113.3
Rater 2	2.2	1633.9	131.2	248.1	52.6
Rater 3	0.4	1298.2	203.3	260.8	110.3
Rater 4	5.3	448.5	86.8	72.4	64.5

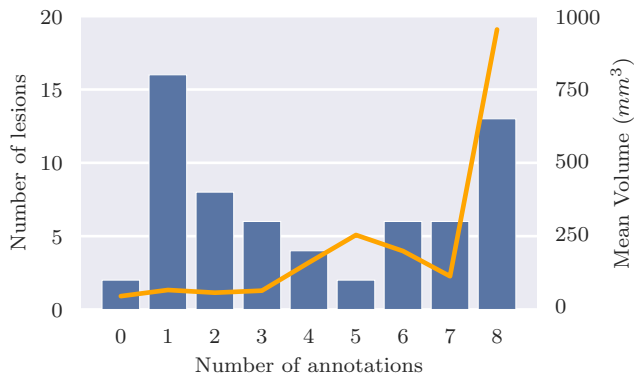

Fig. 5. Rater agreement and lesion volume. The blue bars show how many GT lesions were detected by different numbers of annotations (four raters with two sessions each). A GT lesion is counted as detected if the annotation has segmented over 10% of its voxels. For each level of rater agreement, the mean volume of lesions is shown in orange. For example, 13 GT lesions were present in all eight annotations, and these lesions had a mean volume of 956mm^3.

lesions were detected by all four raters and in both sessions, suggesting obvious lesions on which all raters agree, although there may still be disagreement on the exact delineation. On the other side of the figure, 16 GT lesions were detected in only one annotation, meaning that even the same rater did not detect them across sessions. Moreover, two lesions in the senior expert GT were not present in any of the raters' annotations; these two lesions were detected by the automated method. Importantly, 34 (54%) of the 63 GT lesions are present in only one to four annotations. Therefore, if a majority vote was used to create a consensus GT from these eight annotations, more than half of the valid lesions would be removed. Lesions with a higher level of rater agreement tended to have higher volumes, as the mean volume of lesions detected in all eight annotations was 956.0mm^3 compared to 53.1mm^3 for those present in only one annotation.

B. Performance against Ground Truth

Most of the voxels predicted to be lesion voxels by the raters and model were validated by the senior expert, as seen in the voxel-wise precision distribution in Fig. 6. In particular, for 7 of the 9 MR volumes with lesions, the automated segmentation method has a voxel-wise precision of over 98%. On the other hand, the voxel-wise sensitivity scores are significantly lower. Rater 1 has the highest sensitivity, followed by the other three raters who all have median voxel-wise sensitivity between 20-30%. Finally, the median voxel-wise sensitivity for the model is 14%. The patterns of Dice score across raters are similar to voxel-wise sensitivity, albeit slightly higher because of the influence of voxel-wise precision on the Dice similarity.

Lesion-wise sensitivity is derived from voxel-wise sensitivity (see Sec. II-F), so we might expect the distributions of the two metrics to be similar. However, different patterns arise in the distribution of lesion-wise sensitivity. For example, rater 1 has a higher voxel-wise sensitivity and Dice coefficient in general, but this is not reflected in a generally higher lesion-wise sensitivity. This suggests that rater 1 is not detecting more lesions than the other raters, but that rater 1 segments comparatively larger areas for the lesions they detect, which also aligns with the segmented lesion volume statistics presented in Sec. III-A2.

A further difference when comparing voxel-wise and lesion-wise sensitivity is that the automated method has a more similar distribution to the four raters in lesion-wise sensitivity, with its median sensitivity even surpassing that of rater 3. The different patterns in voxel-wise and lesion-wise sensitivity demonstrate the added value of lesion-wise metrics in addition to voxel-wise metrics.

The results in Fig. 6 present an important overall picture. Most of the lesions that were segmented by the raters were accepted by the senior expert in the adjudication phase, which can be seen by the high precision. However, many lesions were missed by each rater, leading to an underestimate of disease involvement. In particular, none of the raters achieved a median lesion-wise sensitivity of 60%.

A comparison of the lesion-wise sensitivities between the two manual segmentation sessions (Fig. 7) indicates that although some intra-expert differences exist (e.g., for rater 2), the differences between experts are more substantial. This aligns with the observations of variability in lesion counts and volumes in Sec. III-A.

Finally, specificity was calculated with respect to the MR volumes. According to the senior expert GT, there were no lesions in three of the 12 MR volumes in this study. Subject 2 had no GT lesions in either the upper (2-U) nor lower (2-L) volumes, whereas subject 1 had no lesions in the upper (1-U) spinal cord, but had two lesions in the lower (1-L) spinal cord. Indeed, raters 3 and 4 correctly segmented no lesions in the three volumes with no GT lesions. Rater 1, however, segmented one lesion in volume 2-U in both sessions. Rater 2 segmented one lesion in 2-L in both sessions, one lesion in 2-U in the first session, and one lesion in 1-U in the second session. The model segmented five lesions in 2-U, but returned no lesions in 2-L nor 1-U. In other words, raters 1 to 4 had volume-level specificities of 66.7%, 33.3%, 100%, and 100%, respectively, while the model had a specificity of 66.7%.

C. Potential Added Value of the Automated Method

To simulate the potential added value of the model to clinical practice, the number of lesions detected by the automated method but not by individual raters was calculated and is presented in Table II. A mean number of 4.9 (min=3, max=8) lesions could have been potentially retained with the help of the model, which could lead to a mean lesion-wise sensitivity increase of 7.8% (min=4.8%, max=12.7%) for each rater.

D. Evaluation of Automated Consensuses

Using a majority vote on the nine segmentation masks per MR volume (one automated and eight manual segmentations) returned a total of 39 lesions with a mean volume of 112.1mm^3. A STAPLE consensus removed fewer of the

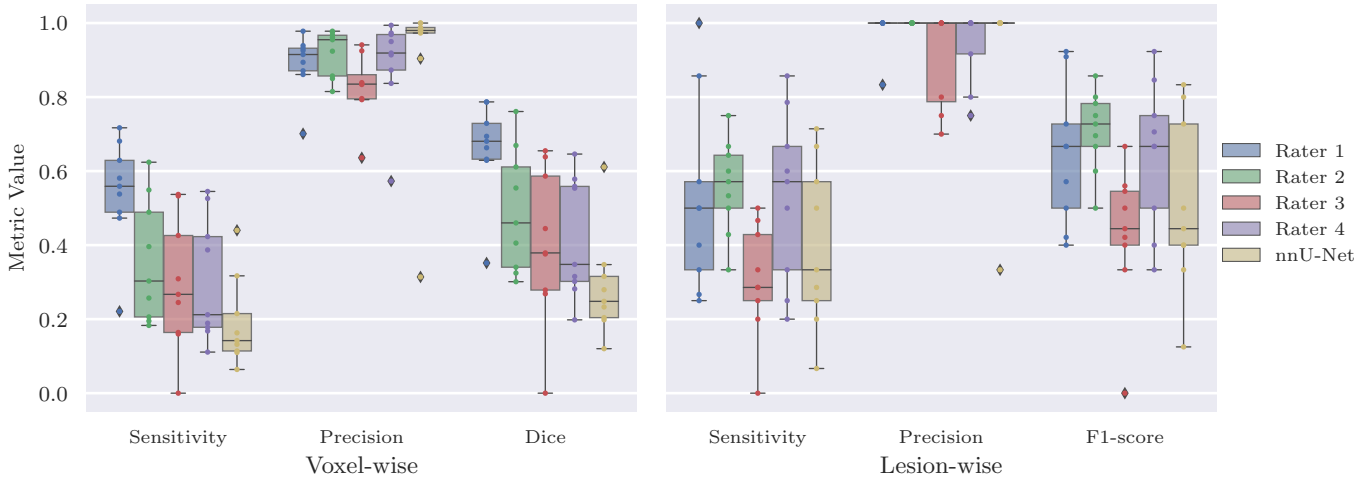

Fig. 6. Lesion-wise and voxel-wise metrics for each rater and the algorithm predictions in the first annotation session evaluated against the senior expert GT. Metrics are first calculated for each of the nine MR volumes with GT lesions (individual points), and boxplot distributions over these nine values are shown.

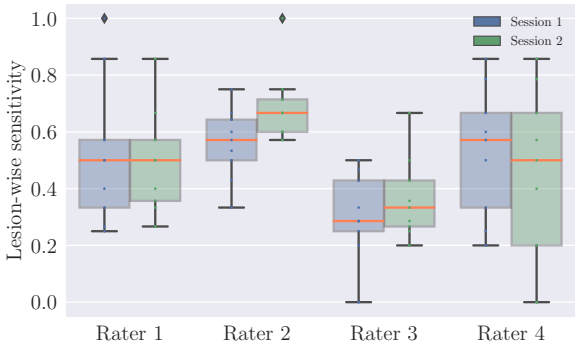

Fig. 7. Comparison across the two manual segmentation sessions of lesion-wise sensitivity, which is evaluated against the senior expert GT.

TABLE II
NUMBER OF GT LESIONS DETECTED BY THE NNU-NET MODEL AND MISSED BY A RATER, FOR BOTH SEGMENTATION SESSIONS.

Rater	Session 1	Session 2
Rater 1	3	3
Rater 2	7	5
Rater 3	8	5
Rater 4	3	5

lesions, leaving 59 lesions with a mean volume of 216.0mm^3. A total of 75 lesions with a mean volume of 272.1mm^3 were present after taking a union of all nine masks, i.e., accepting any voxel marked in any of the nine masks as a lesion voxel (*Any Rater*).

We calculated the same evaluation metrics for the consensus segmentations as before for the individual rater segmentations, comparing each consensus to the GT adjudicated by a senior expert. These results are presented in Table III. As might be

TABLE III
EVALUATION OF CONSENSUS METHODS AGAINST SENIOR EXPERT ADJUDICATED GT. SHOWN ARE THE MEDIAN SCORES ACROSS THE NINE MR VOLUMES WITH GT LESIONS. THE MEDIAN RATER PERFORMANCE IS GIVEN FOR COMPARISON. *(Sens=Sensitivity, Prec=Precision)*

	Voxel-wise			Lesion-wise		
Consensus	Sens	Prec	Dice	Sens	Prec	F1
Majority	28.2	**97.9**	44.0	40.0	100.0	57.1
STAPLE	63.2	84.4	74.5	60.0	100.0	75.0
Any Rater	**99.3**	72.3	**83.7**	**100.0**	100.0	**100.0**
Raters	33.1	88.9	48.9	50.0	100.0	66.7

expected after considering the high precision of individual raters observed in Sec. III-B, the '*Any Rater*' consensus best approximated the final senior expert GT, with a median Dice similarity of 83.7%. Although the majority consensus was highly precise, with a median voxel-wise precision of 97.9%, it discarded many valid lesion voxels, resulting in a low voxel-wise sensitivity of 28.2% and Dice of 44.0%. Finally, STAPLE kept more valid lesion voxels, with a median voxel-wise sensitivity of 63.2% and Dice of 74.5%. As a result, STAPLE outperformed the median rater in voxel-wise Dice and lesion-wise F1-score, whereas majority voting did not.

The previous analysis focused on how well automated consensuses approximated the senior expert GT, but we also examine how reported performance metrics would change by using an automated consensus as GT. For this, we calculated the median Dice scores by taking the median first over the nine MR volumes and then over the raters and sessions. The median Dice score for the raters against the senior expert GT was 48.9%, but when evaluated against the STAPLE and majority consensuses, this rose to 51.8% and 67.3%, respectively, whereas reported performance would drop to 40.2% if the '*Any Rater*' mask was used as GT. The median Dice of the model changes in a similar way, with scores of 16.4%, 24.8%,

26.4% and 42.3% using the '*Any Rater*', adjudicated, STAPLE or majority consensuses, respectively, as GT.

IV. Discussion

A. Inter-rater Variability

Considerable inter-rater variability was observed in this study. The total number of lesions segmented by each rater in all 12 MR volumes ranged from 28 to 60, and only 13 of the 63 lesions validated by the senior expert were segmented by all four raters in both sessions. Even when raters agreed on lesions, the segmented lesion volumes varied significantly. For example, the median lesion volumes segmented by raters 1 and 3 were over double that of rater 2, and the maximum lesion volume segmented by rater 1 (2549mm^3) was over five times that of rater 4 (449mm^3). This inter-rater variability leads to very different depictions of the total lesion load and the disease involvement for individual patients.

Clinical patient stratification based on the number of spinal cord lesions can inform patient care and treatment decisions. The disagreements observed between raters, and between the two segmentations of the same rater, could therefore lead to different patient care pathways. For instance, for subject 1, rater 3 segmented no lesions in one session, whereas all other manual annotations showed a lesion of approximately 100mm^3 for this subject. This single lesion could change the prognosis for the patient and may thus affect therapeutic decisions.

The raters were instructed to segment lesions only if they had a high confidence that they were indeed lesions. This may have led to increased variability, as it is not just a question of detection of hyper-intense areas, but also a subjective assessment of the probability those areas represented lesions. Furthermore, the raters were instructed that segmentation boundaries should include only hyper-intense voxels. However, areas of hyper-intensity can also arise from artifacts and partial volume effects with cerebrospinal fluid, for example. Moreover, what appears as hyper-intense is influenced by the manual adjustment of image contrast by each rater using the ITK-SNAP viewer, which introduces further subjectivity. These factors partly explain the observed disagreements on lesions and lesion volumes.

B. Rater Performance

Evaluated against a ground truth (GT) of a senior expert adjudicating on all possible lesion voxels, both the raters and the automated segmentation method tended to be "precise". Specifically, when the raters or model found and delineated a lesion, the senior expert tended to agree that a lesion existed in that location. On the other hand, the median lesion-wise sensitivity of each rater ranged from 28% to 60%, which demonstrates that detecting lesions in spinal cord MR volumes remains a challenging task and can lead to underestimates of disease activity for patients.

In this study, we considered the segmentations of the senior expert to be the "ground truth" and a lower lesion sensitivity was interpreted to mean that some lesions were missed. However, it is possible that the raters detected and examined the same hyper-intense regions but disagreed as to whether they represented lesions. Indeed, for certain lesions it is challenging to conclusively determine if it is truly a lesion or not based on a single spinal cord MR volume.

Comparing to previous studies, Combès *et al.* [19] reported a Fleiss kappa coefficient of 0.47 in an inter-rater study of new lesion segmentation in brain MR volumes, compared to 0.26 and 0.15 for the two annotation sessions here, indicating a lower agreement between raters in the current study. The same study reported a range of lesion-wise sensitivity of the three raters between 60-66%. The median Dice score was 48.9% in the current study, which is lower than other studies on MS lesion segmentation in brain and spinal cord MR volumes, albeit with differing methodologies. Gros *et al.* [5] reported a median Dice of 60.7% comparing seven raters to a majority consensus in spinal cord MR volumes, Carass *et al.* [9] found a Dice of 66-67% in brain MR volumes, comparing two raters to a STAPLE consensus of the raters and 14 algorithms, while Egger *et al.* [7] observed a mean Dice of 66% over pairwise comparisons between three raters on brain MR volumes. However, caution must be used when comparing these figures, as if a majority vote had been used in the current study in place of an adjudication by a senior expert, the reported median Dice would rise from 48.9% to 67.3%. Moreover, with the small sample size used in this study and by Gros *et al.* [5] (ten MR volumes), these figures may not be representative of the performance over a wider population.

C. Model Performance

The variability among raters and the potential of missing lesions and underestimating disease activity is a motivating factor for the creation and adoption of automated methods as aids to radiologists and neurologists. Indeed, the automated method tested in this study could have increased sensitivity by between 4.8 and 12.7 percentage points. However, taking the union of the segmentations of a clinician and a model is not how an automated method would be used in practice, so future work will conduct a more in-depth study of the usefulness of an automated method in clinical practice, similar to [19].

The high precision and low sensitivity of the algorithm indicate that a different threshold on the model outputs may yield a better balance between sensitivity and precision. Model output probability maps were thresholded at the default value of 0.5. However, we observed a low model output score ($<$ 0.01) for the majority of voxels labelled as lesions in the GT, so adjusting the binarisation threshold could increase sensitivity, as found in previous studies [19], [20]. The default value of 0.5 was kept for this study as the segmentation masks provided to the senior expert to create an adjudicated GT had been binarised using this threshold (see Sec. II-D1).

The high precision and low sensitivity of the model could also be influenced by differences between the GT masks used for training the model and those used for evaluation in the current study. In both cases, the masks were validated by a senior expert; however, the initial mask for evaluation was based on the union of eight manual segmentations and the

model predictions, whereas for training each mask was created by just one expert. As a result, it is possible that not all lesions were detected in the segmentation masks for training, leading the model to reproduce a lower lesion sensitivity. The high precision and low sensitivity of individual experts in the current study can inform future choices when training deep learning models to segment spinal cord lesions. Moreover, the high variability observed demonstrates that the choice of GT used to evaluate a model could have a substantial effect on reported results.

D. Automated Consensuses

A consensus created with a voxel-wise majority vote poorly approximated the adjudication of the senior expert, with a Dice similarity of 44.0% when taking the median over the nine MR volumes containing GT lesions. STAPLE improved on this approximation with a median Dice of 74.5%, while the best approximation of the GT in this study was given by a union of all the manual segmentations and model output, giving a median Dice of 83.7%. This result is influenced by the high precision of both individual raters and the model; majority voting may still be useful in scenarios where raters are expected a priori to have a high sensitivity and low precision.

E. Limitations and Future Work

Further analysis did not indicate significant differences in rater performance where both STIR and T2-w acquisitions were available compared to T2-w alone. However, the added value of STIR to lesion sensitivity has been discussed in the literature [21], and it is one of the recommended sequences for spinal cord imaging [22]. We note that no STIR volumes were used by the model, which could be another factor partly explaining the lower voxel-wise sensitivity relative to the raters. Future work will use both modalities as the inputs of an automated segmentation method, and assess the potential of such a method to further aid spinal cord MS lesion detection in clinical practice.

Future work will also include exploring a better balance between sensitivity and precision in the nnU-Net model, and assessing the impact of the definition of lesion-wise metrics. For example, Gros *et al.* [5] determined a GT spinal cord lesion to be correctly detected if 25% of its voxels had been segmented, whereas we opted for the lower threshold of 10% used by Commowick *et al.* [18] for brain lesions. This decision was motivated by the large variability observed between raters in volumes of segmented lesions. However, this threshold and the metrics in general need to be assessed as to whether they are clinically meaningful.

A limitation of this study is the small sample size of six patients. We make no claim that the reported statistics represent the true sensitivity and precision of experts in clinical practice, nor even for the specific experts in this study. Moreover, while we have demonstrated that significant variability can exist between raters, more large-scale studies are required to determine the factors that influence this variability.

REFERENCES

[1] A. Kutzelnigg and H. Lassmann, "Chapter 2 - Pathology of multiple sclerosis and related inflammatory demyelinating diseases," in *Handbook of Clinical Neurology*, ser. Multiple Sclerosis and Related Disorders, D. S. Goodin, Ed. Elsevier, Jan. 2014, vol. 122, pp. 15–58.

[2] A. J. Thompson *et al.*, "Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria," *The Lancet Neurology*, vol. 17, no. 2, pp. 162–173, Feb. 2018.

[3] S. Leguy, B. Combès, E. Bannier, and A. Kerbrat, "Prognostic value of spinal cord MRI in multiple sclerosis patients," *Revue Neurologique*, vol. 177, no. 5, pp. 571–581, May 2021.

[4] H. Kearney, D. H. Miller, and O. Ciccarelli, "Spinal cord MRI in multiple sclerosis—diagnostic, prognostic and clinical value," *Nature Reviews Neurology*, vol. 11, no. 6, pp. 327–338, 2015.

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

[6] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101759, Oct. 2020.

[7] C. Egger *et al.*, "MRI FLAIR lesion segmentation in multiple sclerosis: Does automated segmentation hold up with manual annotation?" *NeuroImage: Clinical*, vol. 13, pp. 264–270, Jan. 2017.

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

[9] ——, "Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis," *Scientific Reports*, vol. 10, no. 1, p. 8242, May 2020.

[10] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[11] S. Warfield, K. Zou, and W. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[12] S. Vukusic *et al.*, "Observatoire Français de la Sclérose en Plaques (OFSEP): A unique multimodal nationwide MS registry in France," *Multiple Sclerosis Journal*, vol. 26, no. 1, pp. 118–122, 2020.

[13] G. Krinsky, N. M. Rofsky, and J. C. Weinreb, "Nonspecificity of short inversion time inversion recovery (STIR) as a technique of fat suppression: pitfalls in image interpretation." *AJR. American journal of roentgenology*, vol. 166, no. 3, pp. 523–526, Mar. 1996.

[14] P. A. Yushkevich *et al.*, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, Jul. 2006.

[15] M. Antonelli *et al.*, "The Medical Segmentation Decathlon," *Nature Communications*, vol. 13, no. 1, p. 4128, Jul. 2022.

[16] B. De Leener *et al.*, "SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data." *NeuroImage*, vol. 145, no. Pt A, pp. 24–43, Jan. 2017.

[17] T. A. Lampert, A. Stumpf, and P. Gancarski, "An Empirical Study Into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation," *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, vol. 25, no. 6, pp. 2557–2572, Jun. 2016.

[18] O. Commowick *et al.*, "Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure," *Scientific Reports*, vol. 8, no. 1, p. 13650, Sep. 2018.

[19] B. Combès *et al.*, "A Clinically-Compatible Workflow for Computer-Aided Assessment of Brain Disease Activity in Multiple Sclerosis Patients," *Frontiers in Medicine*, vol. 8, p. 740248, 2021.

[20] B. R. Hussein *et al.*, "A study on loss functions and decision thresholds for the segmentation of multiple sclerosis lesions on spinal cord MRI," Nov. 2022.

[21] N. B. Nayak, R. Salah, J. C. Huang, and G. M. Hathout, "A comparison of sagittal short T1 inversion recovery and T2-weighted FSE sequences for detection of multiple sclerosis spinal cord lesions," *Acta Neurologica Scandinavica*, vol. 129, no. 3, pp. 198–203, Aug. 2013.

[22] M. P. Wattjes *et al.*, "2021 MAGNIMS–CMSC–NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis," *The Lancet Neurology*, vol. 20, no. 8, pp. 653–670, Aug. 2021.