Towards Deep Learning-assisted Quantification of Inflammation in Spondyloarthritis: Intensity-based Lesion Segmentation

Carolyna Hepburn¹, Hui Zhang¹, Juan Eugenio Iglesias¹,²,³, Alexis Jones⁴, Alan Bainbridge⁵, Timothy J.P. Bray⁵∗ and Margaret A. Hall-Craggs⁵

¹Centre for Medical Image Computing, Department of Computer Science, University College London, UK
²Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, USA
³Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Boston, USA
⁴Department of Rheumatology, University College London Hospitals NHS Foundation Trust, London, UK
⁵Centre for Medical Imaging, University College London, London, UK

Abstract

Purpose: To develop a semi-automated, AI-assisted workflow for segmentation of inflammatory lesions on STIR MRI of sacroiliac joints (SIJs) in adult patients with axial spondyloarthritis.

Methods: Baseline human performance in manual segmentation of inflammatory lesions was first established in eight patients with axial spondyloarthritis recruited within a prospective study conducted between April 2018 and July 2019. To improve readers’ consistency a semi-automated procedure was developed, comprising (1) manual segmentation of ‘normal bone’ and ‘disease’ regions (2) automatic segmentation of lesions, i.e., voxels in the disease region with outlying intensity with respect to the normal bone, and (3) human intervention to remove erroneously segmented areas. Segmentation of disease region (subchondral bone) was automated via supervised deep learning; 200 image slices (eight subjects) were used for algorithm training with cross validation, 48 (two subjects) – for testing and 500 (20 subjects) – for evaluation based on visual assessment. The data, code, and model are available at https://github.com/c-hepburn/Bone_MRI.

Human and model performance were assessed in terms of Dice coefficient.

Results: Intra-reader median Dice coefficients, evaluated from comparison of manual segmentation trials of inflammatory lesions, were 0.63 and 0.69 for the two readers, respectively. Inter-reader median Dice was in the range of 0.53 to 0.56 and increased to 0.84 using the semi-automated approach. Deep learning model ensemble showed average Dice of 0.94 in subchondral bone segmentation.

Conclusions: We describe a semi-automated, AI-assisted workflow which improves the objectivity and consistency of radiological segmentation of inflammatory load in SIJs.

1 Introduction

Axial spondyloarthritis (axSpA) is a rheumatic disease that primarily affects axial joints of the spine and causes pain, stiffness, and disability [1]. A hallmark of early disease is inflammation in the subchondral bone marrow of sacroiliac joints, referred to as bone marrow edema (BME) [2]. In current clinical practice inflammation is typically detected with short tau inversion recovery (STIR) magnetic resonance imaging (MRI) [3]. Accurate and precise identification and quantification of inflammation is necessary for therapeutic decision making to facilitate effective treatment, control pain and prevent disability. At present, there are several methods for assessing inflammation on STIR MRI and these are based on visual assessment. In clinical practice, the extent of inflammation is described verbally with no numerical metric to enable quantification. Clinical research employs semi-quantitative scoring methods of BME [4], [5], [6]; however there are limitations to this approach. These methods evaluate only a limited number of slices, force the observer to make binary decisions and are laborious. Furthermore, visual assessment is a predominantly intensity-based approach: “the stronger the hyperintense signal the more likely it reflects active inflammation” [7]. Absence of the intensity threshold between normal and abnormal bone makes it difficult to define subtle lesions or boundaries. This impacts on whether lesions meet the minimal areas to be determined as abnormal [2] and causes variability in STIR MRI interpretation. One approach to addressing the quantification problem is to segment inflammatory lesions and thus estimate inflammatory load. The use of thresholding as a segmentation method can assist the observer with intensity-based judgments, reducing subjectivity in image interpretation [8]. A further benefit of segmentation is that lesion location is specified. Recently proposed semi-automated methods for detecting and quantifying only BME also addressed these issues [9], [10]. However, these approaches require the user either to identify lesion location - with the potential that some lesions might be missed - or to manually segment bone, which is time consuming and thus

*Joint senior authorship
impractical in clinical practice. The drawback of manual segmentation of specific areas of bone [10] was addressed in a recent study, which also aimed to demonstrate computational feasibility of a fully automated segmentation of BME [11]. Ultimately, a tool that can automatically detect and quantify inflammation in a reproducible, fast manner is desirable. With advances in artificial intelligence, it is natural to consider supervised deep learning to automate segmentation [12]. Given the challenges in inflammation assessment, the goal of this study was to investigate the reliability of segmentation as means to provide a robust reference standard for algorithm training. As a result, a semi-automated, AI-assisted workflow to improve observer’s precision and reduce workload is proposed.

2 Materials and Methods

2.1 Data

Data were taken from a completed prospective longitudinal study (16 females, 14 males; mean age 42.7 years). The study was conducted at University College London Hospital between April 2018 and July 2019 with the aim of evaluating responsiveness and response prediction using quantitative imaging biomarkers and performed with institutional review board approval (REC reference 15/LO/1475). Patients diagnosed with axSpA according to 2009 ASAS criteria [7] underwent pre- and post-biologic therapy MRI scans of SIJs (n=30). Patients were excluded if they had a contraindication to MRI scanning. The protocol consisted of STIR and T1-weighted TSE sequences, acquired in an oblique coronal plane with the same FOV on a 3T Philips Ingenia scanner (Philips Healthcare, Best, Netherlands). MRI parameters are detailed below.

- STIR: repetition time (TR) 5316ms, echo time (TE) 60ms, inversion time 210ms, echo train length (ETL) 21, flip angle 90deg, slice thickness 3mm, spacing between slices 3.3mm, pixel spacing 0.59 x 0.59 (mm), image matrix 336 x 336, number of slices 23-25

- T1-weighted TSE: TR 1012ms, TE 8ms, ETL 3, flip angle 90deg, slice thickness 3mm, spacing between slices 3.3mm, pixel spacing 0.59 x 0.59 (mm), image matrix 336 x 336, number of slices 23-25

2.2 Manual and Semi-automated Segmentation of Inflammatory Lesions on STIR MRI

Human performance in manual segmentation of inflammatory lesions from two readers (SupMat A, B) was evaluated in a quantitative manner (section 3.1). To mitigate variability a procedure was developed, comprising the following steps (see Figure 1 for study design overview):

- Manual segmentation of interforaminal sacral bone, which is typically spared from inflammation [5], referred to as normal bone (SupMat B)

- Estimation of two thresholds from the normal bone intensity distribution, allowing for a possible range of inflammatory loads and less abrupt lesion boundaries (SupMat C)

- Manual segmentation of the subcortical bone (iliac and sacral) of the fibrocartilaginous (anterior) portion of the SIJs (SupMat D), the most inflamed bone in axSpA, referred to as the disease region. Segmentation was initially performed for three subjects on T1W images, which were used to pre-train a deep learning algorithm to facilitate further segmentation. The algorithm was subsequently re-trained once all segmentations had been corrected (section 2.3, section 3.3)

- Thresholding within the disease region: voxels in the mask were assigned labels of 0, 1 and 2 if voxel intensity at the corresponding location in the STIR image was below, equal, or above lower and upper intensity limit.

- Automatic removal of erroneously segmented regions, containing < 4 pixels (≈ 1.39mm²), on each slice, due to noise and small vessels

- Manual correction of the final segmentation by a human observer based on morphology and anatomical location (SupMat D)
2.3 Supervised Deep Learning Segmentation of Disease Region on T1W MRI

248 T1W image slices were used for model training and testing. The reference standard for algorithm training is described in the SupMat 13. Data was partitioned at subject level at random into two sets: 200 image slices for training with four-fold cross validation 14 and 48 image slices for testing. Further data partition into subsets for training and folds for validation was performed at subject level to avoid correlation between image slices, used to train and validate a model. The final model averaging ensemble was further evaluated, based on visual assessment, on the remaining 500 image slices (20 subjects) (Figure 1).

2.4 Model Training

A convolutional neural network with two-dimensional U-Net architecture 15 (SupMat E) was trained on mini batches by optimizing binary cross entropy loss 16 using the Adam optimizer 17. The architecture included batch normalization to keep the distribution of convolution layers outputs’ fixed, allowing faster convergence 15. The network was trained with pre-processed data augmented on the fly, which allowed a substantial increase of the diversity in the training samples (SupMat F). To keep the number of network parameters updates independent of the data set size, 350 augmentation steps were performed at each training epoch 19. At each step, a random batch was selected from the available pool, augmented, and fed into the model. Data shuffling ensured that the same batch contained different image slices every epoch. To reduce individual models’ errors, the optimal model was re-trained three times and performance of the ensemble was evaluated on the test data. Finally, the model was re-trained three times using all the data and evaluated further on the remaining 40 scans from 20 subjects. A publicly available implementation of the U-Net in Pytorch was used (https://github.com/jvanvugt/pytorch-unet).

2.5 Evaluation Metrics

The similarity of a pair of binary segmentations was evaluated with the Dice coefficient, defined for the class of interest (abnormal or background) as the ratio of number of pixels (voxels) having identical location in both segmentations to the average of number of pixels (voxels) in each segmentation 20:

\[ Dice = \frac{2|S_1 \cap S_2|}{|S_1| + |S_2|} \]  

\( S_{i, i \in \{1,2\}} \) represents a point set, containing pixel (voxel) coordinates, subscript refers to the first or second segmentation and \( S_1 \cap S_2 \) is the intersection of the sets. We refer to the Dice coefficient as the area or volume overlap, depending on whether two segmented areas or volumes were compared. To evaluate model performance during training, a differentiable generalization of Dice was used, implemented as 21:

\[ Dice = \frac{2 \sum_{i=1}^{N} r_i p_i}{\sum_{i=1}^{N} r_i^2 + \sum_{i=1}^{N} p_i^2} \]
where summation runs over pixels of reference standard, \( r_i \in R \) and network probability map, \( p_i \in P \). The similarity between two binary segmentation volumes is defined in terms of absolute volume difference divided by the sum of the compared volumes [20]:

\[
V_S = 1 - \frac{|S_1| - |S_2|}{|S_1| + |S_2|}
\]

3 Results

3.1 Manual vs Semi-automated Segmentation of Inflammatory Lesions: Intra/Inter-Reader Variability

Figure 2 shows the 3D rendering of some manual segmentation trials of inflammatory lesions from both readers. It demonstrates qualitatively differences in lesion location, degree of fragmentation and load between trials. Figure 3 shows the sum of manual trials, indicating the difficulties in image interpretation, namely, subtle inflammation and unclear lesion boundaries, and its’ semi-automated counterpart. Examples of automatic segmentation via thresholding with no correction by readers are presented in the SupMat [C]. Readers’ self-consistency and agreement, evaluated in terms of volume overlap, are shown in Figure 4a, b. Overlap values ranged from 0.28 to 0.87. Intra- and inter-reader median overlap values were 0.63 and 0.69 for reader 1 and 2 and in the range 0.53-0.56, respectively. There were occasions when the volumes from manual trials were dissimilar, affecting Dice score (Figure 4d) and very similar, however, readers disagreed on lesion location (SupMat [I]). Metric values depended on the trials compared. For the semi-automated pipeline, the between-reader median volume overlap was 0.84, representing an increase of 28-31% compared to pure manual segmentation (Figure 4c). There was one outlier where the agreement was reduced; review of the images indicated that the disagreement mostly related to the presence of inflammation in the joint space, where blood vessels can be misinterpreted (SupMat [I]). There was a time gain for the semi-automated pipeline of 13.34min on average for reader 1 and 3.12min on average for reader 2 compared to pure manual delineation. Visual scoring of BME (SPARCC method [4]), using automatic segmentations with removed residues on pre-defined slices took 3 min on average for reader 1 (SupMat [J]).
Figure 3: STIR MRI slices in oblique coronal plane for the same subjects as in the figure 2 (top to bottom). Sum of manual segmentation trials (second column) indicates pixels that were labelled as abnormal once (dark blue), twice (light blue), trice (yellow) and four times (red). Corrected automatic segmentations by the two readers (column 3 and 4) with pixels, segmented at lower (yellow) and both (red) thresholds.
Figure 4: Volume overlap evaluated from comparison of different manual segmentation trials of the same reader for each reader (a), different readers (b) and corrected automatic segmentations (thresholding at lower intensity limit), (c). \( R_{ij} \) stands for reader with the first subscript corresponding to the reader and second - to manual segmentation trial, when applicable. Scatter plot of overlap vs volume similarity from comparison of some trials from both readers (d). Vertical line corresponds to similarity value when one volume, \( V \) is twice larger than the other, \( V^* \).
3.2 Semi-automated Segmentation Sensitivity to Change in Inflammatory Load

Figure 5 shows automatic segmentations corrected by the second reader on pre- and post-biologic treatment scans for one subject. Possible inflammatory volume ranges are $(106.3, 115.3) \times 10^3 \text{ mm}^3$ and $(31.3, 46.3) \times 10^3 \text{ mm}^3$ for pre- and post-treatment scans respectively, indicating a reduction in inflammation. Regions of acute inflammation showed a reduction in extent and intensity after treatment, although there is a persistent, slight hyperintensity compared to the normal interforaminal bone, which manifests as an increase in the proportion of inflammation captured by the lower of the two thresholds.

Figure 5: STIR MRI slices in oblique coronal plane for the same subject (left), pre-biological treatment (top) and post-biological treatment MRI scan (bottom). Super-imposed corrected by the second reader automatic segmentations (middle), as well as 3D rendering (right) qualitatively show reduction in inflammation.
3.3 Automatic Segmentation of Disease Region

Optimal hyper-parameters were searched with cross validation procedure by varying network depth, number of kernels per convolution layer, kernel size and number of epochs. A batch size of four and a learning rate of 0.001 were kept constant. Training was stopped when average area overlap (section 2.5), evaluated on training data, reached a plateau (Figure 6a). The optimal model did not overfit; however, there were fluctuations in its performance on the fourth validation fold (Figure 6b). This may be attributed to the fact that one scan was acquired with different angulation with respect to the other scans. The optimal network was retrained three times using all training data (200 T1W image slices). From each run a model yielding the highest average area overlap was chosen. Model averaging ensemble yielded 0.94 average area overlap on the test data, ranging from 0.85 to 0.98 (Figure 6c). Examples of automatically segmented disease region and corresponding reference standard are shown in the SupMat K. Model averaging ensemble was further evaluated based on visual assessment by the second reader, using the remaining 20 subjects (40 scans). Model failed or its performance was worse for three subjects with very abnormal bone, i.e., very high fat content or strong sclerosis; such features were absent in the training data. For the rest of the scans, model performance was either perfect or subject to minor corrections (SupMat K). Consequently, three subjects with very abnormal bone were added to the training data and model was retrained three times for public access.

Figure 6: Mean area overlap vs training epoch for different training data subsets (a) and validation folds (b). Each point represents a score, averaged over classes (foreground & background), samples in a mini batch and 350 augmentation steps. Area overlap from pair-wise comparison of reference standard and rounded prediction on the test data from model averaging ensemble (three runs using all training data) (c).

4 Discussion

We first investigated the reliability of segmentation of inflammatory lesions on qualitative STIR MRI of adult patients with axSpA to provide a robust reference standard for deep learning algorithm training. Based on quantitative evaluation of human performance in manual segmentation we developed a semi-automated procedure to mitigate between-reader variability with the following steps (1) segmentation of subchondral bone region, where inflammation typically occurs, and normal (sacral) bone region, spared from inflammation, (2) automatic segmentation of inflammatory lesions within subchondral bone via two thresholds, determined from the intensity distribution of ‘normal bone’ voxels, and (3) human intervention to remove erroneously segmented areas. We showed that this procedure produced a substantial improvement in segmentation consistency compared to manual segmentation. To ensure the practicality of the approach, we automated segmentation of the subchondral bone region via supervised deep learning. One of the limitations of the proposed approach is the fixed value of the lower threshold. In practice, the observer could adjust the lower intensity limit. This, however, would introduce further variability, that needs to be quantified. Secondly, although the approach mitigates the variability, it does not eliminate it. Therefore, preparation of a high-quality dataset for definitive algorithm training for use in clinical practice requires radiologists’ consensus. Accurate detection and quantification of inflammatory load on STIR MRI is difficult, in large part because there is no defined intensity threshold between normal and abnormal water content resulting from increase in cellularity and/or expansion of the extracellular space [22], [23]. Furthermore, several sources influence the way we interpret the signal: small increases in water content indistinguishable from natural variations in bone marrow composition, confounding factors such as sclerosis and fat metaplasia masking abnormal signal, quality of fat suppression, sensitivity of the receiving coil, noise etc. These give rise to ambiguity in definition of any type of inflammatory lesion visible on STIR MRI. “The stronger the hyperintense signal the more likely it reflects active inflammation (intensity of the hyperintense signal is similar
to that of blood vessels or spinal fluid” [7]. “At present it is not possible to give a more precise definition of the minimum size (area) of the BMO which is necessary for it to be described as ‘positive’” [2]. The reader’s image interpretation relies on a subjective perception of brightness, which is at the origin of variability either in manual segmentation or scoring [21]. The idea of determining an optimal threshold to quantify BME has been addressed in several studies [10, 25]. However, such analysis relied on manual segmentation and as our data suggest, using manual segmentation as a “gold standard” does not necessarily yield correct answers especially in cases when inflammation is subtle or precise lesion boundary cannot be identified due to intensity gradient. Furthermore, a recent study [11] re-established the threshold value developed in earlier work [10], finding an optimal threshold of 1 compared to 1.5 in the prior study. It seems logical that a threshold which depends on reference standard provided by human observers is not desirable. Therefore, a major advantage of the proposed approach is that intensity-based judgements are removed from the observer. The use of an intensity-based thresholds derived from normal marrow means that the choice of voxels is primarily influenced by the physical properties of the tissue (with increased signal in areas of inflammation) and allows for a possible range of inflammatory load. In conclusion, we propose a workflow that could be turned into a tool to improve the objectivity and consistency of radiological assessment of inflammation in the sacroiliac joints and in a long-term goal, prepare a large high-quality dataset needed for deep learning algorithm training to fully automate segmentation procedure with the potential to substantially improve decisions around diagnosis, disease monitoring and treatment.

References

[1] Sieper J, Braun J, Dougados M, Baeten D. Axial spondyloarthritis. Nat Rev Dis Prim. 2015;1(7):1-17
[2] Rudwaleit M, Jurik AG, Hermann KGA, Landewé R, Van Der Heijde D, Baraliakos X, et al. Defining active spondylitis on magnetic resonance imaging (MRI) for classification of axial spondyloarthritis: A consensus approach by the ASAS/OMERACT MRI group. Ann Rheum Dis. 2009;68(10):1520–7
[3] Maksymowycz WP. The role of imaging in the diagnosis and management of axial spondyloarthritis. Nat Rev Rheumatol. 2019;15(11):657–72
[4] Maksymowycz WP, Inman RD, Salonen D, Dhillon SS, Williams M, Stone M, et al. Spondyloarthritis Research Consortium of Canada magnetic resonance imaging index for assessment of sacroiliac joint inflammation in ankyllosing spondylitis. Arthritis Care Res. 2005;53(5):703–9
[5] Maksymowycz WP, Dhillon SS, Chiochwanisawakit P, Pedersen SJ, Martinez B, Östergaard M et al. Development and Validation of Web-based Training Modules for Systematic Evaluation of Active Inflammatory Lesions in the Spine and Sacroiliac Joints in Spondyloarthritis. J Rheumatol. 2009;36:48-57
[6] Navallas M, Ares J, Beltrán B, Lisbona MP, Maymó J, Solano A. Spondylitis associated with axial spondyloarthritis: New concepts and latest trends. Radiographics. 2013;33(3):933–56
[7] Sieper J, Rudwaleit M, Baraliakos X, Brandt J, Braun J, Burgos-Vargas R, et al. The Assessment of SpondyloArthritis International Society (ASAS) handbook: A guide to assess spondyloarthritis. Ann Rheum Dis. 2009;68(SUPPL. 2):ii1-ii44
[8] Mayerhoefer ME, Breitenscher M, Hofmann S, Aigner N, Meizer R, Siedentop H, et al. Computer-assisted quantitative analysis of bone marrow edema of the knee: Initial experience with a new method. Am J Roentgenol. 2004;182(6):1399–403
[9] Zarco P, Almodóvar R, Bueno Á, Molinero LM, Moreno M, Juñola X, et al. Development and validation of SCAISS, a tool for semi-automated quantification of sacroilitis by magnetic resonance in spondyloarthritis. Rheumatol Int. 2018;38(10):1919–26
[10] Kucyba I, Tabor Z, Polak J, Urbanik A, Wojciechowski W. The semi-automated algorithm for the detection of bone marrow oedema lesions in patients with axial spondyloarthritis. Rheumatol Int. 2020;40(4):625–33
[11] Rzecki K, Kucyba I, Gut D, Jarosz A, Nabaglo T, Tabor Z, Wojciechowski W. Fully automated algorithm for the detection of bone marrow oedema lesions in patients with axial spondyloarthritis – Feasibility study. Biocybernetics and Biomedical Engineering 2021;41(2):833-53
[12] Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys. 2019;29(2):102–27
[13] Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 2006;31(3):1116–28
[14] Casella G, Fienberg S, Olkin I. An Introduction to Statistical Learning. Springer Texts in Statistics 2013, 618 p.
[15] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. MICCAI 2015;9351:234–41
[16] Bishop CM. Neural Network for Pattern Recognition. Clarendon Press, Oxford, 1995
[17] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs] 2017
[18] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 32nd Int Conf Mach Learn ICML 2015. 2015;1:448–56

[19] Billot B, Bocchetta M, Todd E, Dalca A V., Rohrer JD, Iglesias JE. Automated segmentation of the hypothalamus and associated subunits in brain MRI. Neuroimage. 2020;223

[20] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging. 2015;15(1)

[21] Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. Proc - 2016 4th Int Conf 3D Vision;565–71

[22] Bollow M, Braun J, Hamm B, Eggens U, Schilling A, König H, et al. Early sacroiliitis in patients with spondyloarthropathy: Evaluation with dynamic gadolinium-enhanced MR imaging. Radiology 1995;194(2):529–36

[23] Bollow M, Fischer T, Reißhauer H, Backhaus M, Sieper J, Hamm B, et al. Quantitative analyses of sacroiliac biopsies in spondyloarthopathies: T cells and macrophages predominate in early and active sacroiliitis - Cellularity correlates with the degree of enhancement detected by magnetic resonance imaging. Ann Rheum Dis. 2000;59(2):135–40

[24] Landewé RBM, Hermann KGA, Van Der Heijde DMFM, Baraliakos X, Jurik AG, Lambert RG, et al. Scoring sacroiliac joints by magnetic resonance imaging. A multiple-reader reliability experiment. J Rheumatol. 2005;32(10):2050–5

[25] Chronaiou I, Thomsen RS, Huuse EM, Euceda LR, Pedersen SJ, Hoff M, et al. Quantifying bone marrow inflammatory edema in the spine and sacroiliac joints with thresholding. BMC Musculoskelet Disord. 2017;18(1):1–8
Supplemental Material

A  Manual Segmentation of Inflammatory Lesions on STIR MRI

Manual delineation and segmentation were performed by a consultant radiologist and a musculoskeletal radiology fellow with over 25 and 5 years of musculoskeletal MRI experience, respectively. 13 subjects with pre- and post-biological treatment scans were selected from the available pool of MRI by the consultant radiologist based on visual assessment to represent a broad range of inflammatory lesions and varying lesion loads. To facilitate and standardize the segmentation, five scans were used to establish guidelines for lesion identification, with suggestions for each of the potential difficulties in interpretation (Figure A.7). Eight scans were used for the study, for which each reader provided two independent segmentation trials. A third non-radiologist reader checked all segmentations for a presence of empty or residual voxels and made necessary adjustments. Readers focused on potential lesions in the subchondral bone marrow abutting the sacroiliac joints, the joint space, and the adjacent lumbar vertebra (L5, mainly seen in the endplates and facet joints). Lesions located above the L4/5 disc were not segmented as this bone was not consistently covered by FOV. We used ITK SNAP software (version 3.8.0) for all segmentations.

| Difficulty | Description | Suggestion |
|------------|-------------|------------|
| No threshold between normal & abnormal bone | Intensity Gradient  
Focal and diffuse oedema are characterized by hyperintense signal, surrounded by intensity gradient. | Comparison with normal marrow, i.e., the midline bone in the sacrum, may assist in determining the abnormality. |
| Subtle Lesions  
Small water content, resulting in slightly increased intensity as compared to the signal from blood vessels and spinal fluid. | Consult corresponding T1W image and segment only if a potential lesion is present on at least 2 consecutive slices. |
| Confounding factors  
Sclerosis and fat metaplasia mask abnormal signal | Consult corresponding T1W image to distinguish between sclerosis and increased fat content |
| No minimal area for a lesion to be assigned as abnormal | Areas of approximately 25 mm² and less | Segment only if a potential abnormality is present on at least 2 consecutive slices (not necessarily at the exact location as 3D lesions are not straight objects) |
| Partial volume effect | Signal arises from different tissues. Therefore, lesion is partially masked. | Do not segment |

Figure A.7: Summary of difficulties, related to manual segmentation, and suggestions to assist diagnostic decision making.

B  Manual Segmentation of Normal Bone on STIR and Disease Region on T1W MRI

Manual segmentation of normal bone and disease region was performed by a non-radiologist reader for ten subjects on STIR and three subjects on T1W images, respectively. A supervised deep learning algorithm was pre-trained with 75 image slices (three subjects) and used to segment disease region on the remaining T1W images from seven subjects. All segmentations were verified and corrected by the two radiologists. The normal bone region was segmented on multiple slices and included the largest possible area of normal interferaminal bone to minimise sampling error and more accurately represent its intensity distribution (Figure B.8). The disease region, i.e., the subcortical bone (iliac and sacral) of the
fibrocartilaginous (anterior) portion of the SIJs, excluded the central sacral bone and foramina, spinal fluid, intervertebral discs, and partially the ligamentous part of the SIJs, i.e., irrelevant regions for inflammation assessment (Figure B.9). The use of the T1W sequence facilitated morphological assessment of the images, especially the facet joints (the anatomy of these regions was more difficult to evaluate on STIR images). In cases when identification of the border between vertebra and discs were difficult, STIR images were consulted and segmentation was adjusted accordingly. T1W and STIR images were acquired in the same plane with the same FOV, enabling usage of disease region mask for consequent segmentation via thresholding on STIR image.

Figure B.8: STIR MRI slice in oblique coronal plane (left) with super-imposed manual segmentation of normal bone region (middle). 3D rendering reveals the segmented structure (right).

Figure B.9: T1W MRI slices in oblique coronal plane for one subject (top) with corresponding super-imposed manual segmentations of disease region (bottom). 3D rendering reveals the segmented structure (right).
C Estimation of two Thresholds from Normal Bone Intensity Distribution

Intensities corresponding to outliers and rare observations of the normal bone intensity distribution are attributed either to noise or increased water content due to natural variation in tissue composition. Therefore, voxels at or above these intensity levels are more likely to be abnormal. Distributions determined from different scans were skewed to varying degrees (Figure C.10a). To standardize the procedure, an upper limit, $L_{\text{upper}}$, was defined as the maximum intensity, $I_{\text{max}}$, of the distribution. A lower limit, $L_{\text{lower}}$, was computed as the sum of the upper quartile, $Q_U$, and a multiple, $n$, of the inter-quartile range, $IQR$, of the distribution: $L_{\text{lower}} = Q_U + n \cdot IQR$. The multiple $n$ was determined automatically as the empirical lower limit value of $Q_U + 1.5 \cdot IQR$ resulted in obvious over-segmentation in certain cases (Figure C.11). Starting with the value of 1.5, the multiple $n$ was incremented by 0.05 until the difference between upper and lower limits was less than the half of the interquartile range, $0 < L_{\text{upper}} - L_{\text{lower}} < IQR/2$. If the condition was initially satisfied, no incrementation was performed.

Figure C.10: Boxplot of normal bone STIR intensity for ten subjects, showing different degree of skewness (a). Histogram of normal bone intensity for one subject (b). Vertical lines represent different threshold values; resulting segmentations are presented in the Figure C.11.

Figure C.11: STIR MRI slice in oblique coronal plane for one subject (left) with super-imposed automatic segmentation via three different thresholds: $Q_U + 1.5 \cdot IQR$ (blue), $Q_U + 2.55 \cdot IQR$ (orange) and maximum intensity (red) of the normal bone intensity distribution (Figure C.10), with $Q_U$ upper quartile and $IQR$ inter-quartile range.
**D Manual Correction of Automatic Segmentation of Inflammatory Lesions Based on Morphology and Anatomical Location**

Regions which were deemed non-inflammatory - for example due to the presence of vessels or artifact - were removed by readers based on morphology and anatomical location. Lesions were either left in place or removed altogether, i.e., the boundaries of lesions were not modified, except when the posterior part of the joint or foramen were segmented along with a potential lesion. Lesions located above the L4/5 disc were removed in consistency with manual segmentations. T1W images were used to assist readers in identification of anatomical structures and regions of increased fat content. The two readers discussed and agreed upon the procedure using automatically segmented images from two subjects (Figure 1, main text).

**E U-Net Architecture**

Network takes as input a mini batch, containing four different T1W image slices, and outputs a batch of corresponding probability maps. Figure E.12 details the operations and illustrates how the size of an individual image slice or a feature map changes, as the slice / map propagates through the network.

![U-Net Architecture Diagram](image)

Figure E.12: Schematic representation of the optimal U-Net architecture. Number of channels and matrix size are, respectively, on the top and in the left corner of a grey shaded box, representing an individual input image slice or a feature map.
Data Pre-processing and Augmentation

Images were normalized by three standard deviation of image intensity distribution, allowing the same intensity scale between subjects and consistency in intensity levels of voxels representing the same tissue across the whole image volume for each subject (Figure F.13). Each pre-processed image and corresponding mask slices undergone elastic deformation (https://github.com/gvtulder/elasticdeform/tree/v0.4.9), affine transformation (rotation, scaling, shearing) and random flip with 0.5 probability, i.e., the order of image array elements along the left-right axis was reversed (Figure F.14). To make the network robust against between-subject variations in the intensity level of bone voxels, intensities were raised to a random power, altering contrast and brightness. All transformation parameters (rotation angle, scaling and shearing factors, power) were randomly sampled from uniform distribution of pre-defined ranges.

Figure F.13: T1W image slice in oblique coronal plane for two subjects: raw (a) and normalized data (b) with corresponding (slice) intensity histograms.
Figure F.14: Original (pre-processed) T1W image slice in oblique coronal plane with corresponding mask (a). Transformed image and mask slices (b). Corresponding (slice) intensity histograms indicate change in intensity levels population after transformation (c).
Examples of Automatic Segmentation of Inflammatory Lesions with no Manual Correction

Figure G.15: STIR image slices in oblique coronal plane without/with super-imposed automatic segmentation for six subjects. Pixels, segmented at lower and both thresholds are marked in yellow and red, respectively.
Examples of Manual Segmentation Trials with Disagreement on Lesion Location

Figure H.16 illustrates specific examples of manual segmentation trials, for which readers agree more on load than location, evaluated in terms of volume overlap and volume similarity.

Figure H.16: STIR image slices in oblique coronal plane with super-imposed manual segmentation trials from both readers for two subjects (left). 3D rendering reveals segmented structures (right). Comparison of these trials yielded volume overlap and similarity values: 0.59 and 0.97 (top), 0.53 and 0.89 (bottom).
I Case with Lower Agreement

Figure I.17: STIR image slices in oblique coronal plane for one subject with super-imposed automatic segmentation, corrected by the two readers. The main source of disagreement is on the potential inflammation in the joint space of SIJs.

J Annotation Time

Following table summarizes time in minutes, taken to manually delineate or segment inflammatory lesions on STIR MRI, time to remove residues on automatic segmentations and consequently, to score BME, using SPARCC approach.

| Scan  | Reader 1 |       |       | Reader 2 |       |
|-------|----------|-------|-------|----------|-------|
|       | Manual delineation only, 1st trial | Residues’ removal | SPARCC scoring | Manual segmentation, 1st trial | Residues’ removal |
| 1003  | 20       | 27    | 2     | 15       | 20    |
| 1007  | -        | 22    | 4     | 10       | 15    |
| 1012  | 18       | 11    | 1     | 25       | 15    |
| 1013  | 56       | 51    | 5     | 25       | 20    |
| 1016  | 60       | 36    | 5     | 20       | 30    |
| 1234  | -        | 20    | 2     | 25       | 15    |
| 3002  | 50       | 19    | 3     | 25       | 15    |
| 4368  | 34       | 14    | 2     | 25       | 15    |
| Average | **39.67 (6 subjects)** | **26.33 (6 subjects)** | **3** | **21.25** | **18.13** |
Examples of Automatic Segmentation of Disease Region

Figure K.18 shows examples of automatically segmented disease region and corresponding reference standard. Figure K.19 and Figure K.20 show automatic segmentation of disease region from seven subjects, for which the reference standard was not available. In cases with very abnormal bone, Figure K.19 model missed some regions. Figure K.20 shows the examples of segmentation nearly at the level of human performance.

Figure K.18: T1W MRI slices in oblique coronal plane for two ‘test’ subjects (top to bottom). Reference standard (middle) and model averaging ensemble (rounded) prediction of disease region (right).
Figure K.19: T1W image slices in oblique coronal plane (top) for three subjects with super-imposed model averaging ensemble rounded prediction (bottom). Subjects exhibit very abnormal bone: high fat content (hyperintense signal; left, middle) and strong sclerosis (hypointense signal; right) were partially missed by the model ensemble.

Figure K.20: T1W image slices in oblique coronal plane (top) for four subjects with super-imposed model averaging ensemble rounded prediction (bottom).