Original Article

Acta Radiologica Open
9(1) 1–10

DOI: 10.1177/2058460120902402
journals.sagepub.com/home/arr
SAGE

Short tau inversion recovery MRI of Modic changes: a reliability study

Per Martin Kristoffersen[1,2], Nils Vetti[1,2], Kjersti Storheim[3,4], Lars Christian Bråten[3,5], Mads Peder Rolfsen[5,6], Jörg Assmus[7] and Ansgar Espeland[1,2]

Abstract

Background: Limited reliability data exist for evaluation of spinal edema changes on magnetic resonance imaging (MRI) with short tau inversion recovery (STIR) sequences.

Purpose: To assess the inter-observer reliability for evaluation of STIR signal increase related to Modic changes (MCs) on MRI of the lumbar spine.

Material and Methods: We prospectively included 120 patients imaged to confirm their eligibility for the AIM (Antibiotics In Modic changes) trial. Three experienced radiologists independently evaluated MCs on T1-/T2-weighted fast spin-echo images and subsequently MC-related STIR signal increases. Inter-observer reliability was analyzed at four endplates (L4–S1) by calculating kappa values and means of differences with 95% limits of agreement.

Results: Overall agreement (mean Fleiss' kappa for all endplates and observers) was very good for presence of STIR signal increase (0.86), and moderate for its categorized height (0.51), anteroposterior extent (0.48), and volume (0.56). For height of region with STIR signal increase measured in % points of vertebral body height, the largest mean of differences was 6.9 and widest range for limits of agreement was ±22.3 for all endplates combined. The corresponding numbers were 11.2 ± 34.5 for anteroposterior extent of the STIR signal increase measured in % points of anteroposterior endplate diameter and 0.9 ± 7.6 for its maximum measured intensity on a % point scale (0% = normal vertebral marrow intensity, 100% = cerebrospinal fluid intensity).

Conclusion: Inter-observer reliability was very good for the presence and intensity of MC-related STIR signal increases, and moderate for their size.

Keywords

Skeletal-axial, magnetic resonance imaging, spine, adults, imaging sequences, observer performance

Received 2 December 2019; accepted 30 December 2019

Introduction

Short tau inversion recovery (STIR) sequences are widely used in magnetic resonance imaging (MRI) to evaluate edematous changes in the skeleton, including the spine. Despite widespread use, limited reliability data exist for spinal evaluations with STIR or other fluid-sensitive fat-suppressed series. Such data were included in articles on spondylarthritis (1–5), fractures (6), Modic changes (MCs) (7), hemangiomas (8), and pedicle screw loosening (9). However, all but one (6) of these reliability studies had only two observers, most (1,3–5,7–9) had small patient samples (n = 25–41), only one (8) included measurements (of signal intensities); the spondylarthritis studies were limited to lesion

[1]Department of Radiology, Haukeland University Hospital, Bergen, Norway
[2]Department of Clinical Medicine, University of Bergen, Bergen, Norway
[3]Research and Communication Unit for Musculoskeletal Health (FORMI), Oslo University Hospital HF, Oslo, Norway
[4]Faculty of Health Science, OsloMet - Oslo Metropolitan University, Oslo, Norway
[5]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway
[6]Division of Orthopaedic Surgery, Oslo University Hospital Ulleval, Oslo, Norway
[7]Competence Centre for Clinical Research, Haukeland University Hospital, Bergen, Norway

Corresponding author:
Per Martin Kristoffersen, Department of Radiology, Haukeland University Hospital, Jonas Liesvei 65, 5021 Bergen, Norway.
Email: per.martin.kristoffersen@helse-bergen.no

detection (1–5). Reliability estimates differed widely in these heterogeneous studies where radiologists and non-radiologists interpreted various fat-suppressed 1.5-T or 3-T series. More comprehensive reliability data are needed for radiologists' lumbar spine evaluations with STIR.

There has been increasing focus on MCs in recent years. MCs are signal changes in the vertebral bone marrow extending from the endplate and are classified into types I (edema type), II (fatty type), and III (sclerotic type) based on T1-weighted (T1W) and T2-weighted (T2W) series (10–12). STIR series are sensitive to edema and are highly relevant for evaluation of MCs. The association between MCs and pain is inconsistent (13–16), but edema type MCs might be symptomatic (17–19). Mechanical, autoimmune, and infectious explanations for MCs have been proposed (20), and various treatments have been and are being tested (21–34). Reliable evaluation of the STIR findings is required to validate their relevance to symptoms and treatment (35–37).

Clinicians and researchers evaluate MCs with a combination of MRI series. The reliability is mostly well described for evaluations with non-fat-suppressed T1W/T2W sequences (38–43), but not for evaluations with fat-suppressed, fluid-sensitive series (7). The primary aim of this study was to assess the inter-observer reliability for evaluation of STIR signal increase related to MCs on MRI of the lumbar spine. For comparison, we also report the inter-observer reliability for the evaluation of these MCs on T1W/T2W fast spin-echo images.

Material and Methods

This reliability study was based on a study-specific MRI of a consecutive subsample (n = 120; 72 women, 48 men; age range = 25–64 years; mean age = 45 years) with chronic low back pain considered for inclusion in the AIM (Antibiotics In Modic changes) trial (32). Inclusion required presence of type I and/or type II MCs at the level of an MRI-confirmed lumbar disc herniation within the preceding two years. All eligibility criteria are listed in the Appendix (Suppl. Table 1). Patients preliminarily eligible for the trial based on these criteria and findings on an existing clinical MRI (n = 220) underwent the new study-specific MRI to confirm or reject their eligibility. All participants included in the study provided written informed consent. The present report adheres to the guidelines for reporting reliability and agreement studies (44).

Images

The 120 study-specific MRI examinations were performed from 15 June 2015 to 2 September 2016 at five centers in Norway, using identical protocols and 1.5-T scanners (Siemens Magnetom Avanto B19). The present study was based on sagittal T1W and T2W fast spin-echo images (=T1/T2) and sagittal STIR images (Table 1).

Evaluation

Three radiologists, all with >10 years of experience in musculoskeletal MRI, independently evaluated the images. The first observer to open the MRI examination saved a mark on the lowest lumbar disc level. All observers reported this level as L5/S1. First, MCs were rated on T1/T2, blinded to other sequences. Later the observers rated STIR findings and decided whether any increased STIR signal was related to an MC visible on T1/T2. The observers were blinded to clinical outcome but knew that patients were preliminarily eligible for the trial. To align their understanding of procedures and rating criteria, the observers rated and discussed MCs and STIR findings in a pilot study (32 MRIs not included in the main study).

On T1/T2, we defined MCs as signal changes in the vertebral bone marrow extending from the endplate, and based rating criteria for MC type and size on prior work (10,11,38,45) (Table 2). Only T1/T2 findings defined MC types I, II and III, not STIR findings. Not recorded as MCs were: (i) changes separated from the endplate; (ii) roundly shaped fatty changes abutting the endplate with a smaller base than height (more likely focal fatty marrow or hemangiomas); and (iii) changes extending through the endplate (Schmorl's hernias).

On STIR, we defined MC-related signal increase as visible increase compared to normal vertebral bone marrow, formed and located as an MC and/or located in or abutting a region with MC on T1/T2 (and not located in a likely hemangioma). MC-related STIR signal increase was evaluated for presence, height, anteroposterior (AP) extent, volume, and maximum intensity (Table 2). STIR signal decrease was not evaluated. STIR signal intensity was measured in the region with most intense MC-related STIR signal, in the cerebrospinal fluid (CSF) and in normal vertebral body marrow (Table 2, Fig. 1). The measurements were made in circular regions of interest available in our PACS with size 25 mm^2 (used for most intense MC-related STIR signal and CSF) and 44 mm^2 (used for normal vertebral body marrow) (Fig. 1). Care was taken to avoid surrounding structures, e.g. intervertebral discs, nerve roots, central vertebral vein. Intensity of CSF varied between levels and was measured at the

Table 1. MRI parameters for sagittal fast spin-echo T1W, T2W, and STIR images of the lumbosacral spine.

Parameter	T1	T2	STIR
TR (ms)	575	3700	5530
TE (ms)	11	87	70
ETL	5	17	20
Acquisitions (n)	2	2	1
Concatenations (n)	2	1	1
Slices (n)	17	17	15
Matrix (frequency × phase)	384 × 269	384 × 269	320 × 224
FOV (mm)	300 × 300	300 × 300	300 × 300
Slice thickness (mm)	4.0	4.0	4.0
Interslice gap (mm)	0.4	0.4	0.4
Voxel size (mm)	1.1 × 0.8 × 4.0	1.1 × 0.8 × 4.0	1.3 × 0.9 × 4.0
Receiver bandwidth (Hz/px)	161	161	182
Phase encoding direction	Head to feet	Head to feet	Head to feet
Saturation pulses	None	Anterior, 30 mm	Anterior, 30 mm
Acquisition time (min:s)	1:48	1:49	1:58
Coverage	From above Th12 to below S2	From above Th12 to below S2	From above Th12 to below S2
Phase oversampling (%)	70	70	70
TI (ms)			160
PAT mode	Grappa	Grappa	None

MRI was performed on 1.5-T Magnetom Avanto scanners (Siemens) with B19 software using integrated spine array coil, but no surface coils.
ETL, echo train length; FOV, field of view; MRI, magnetic resonance imaging; PAT, parallel acquisition technique; STIR, short tau inversion recovery; TE, echo time; TI, inversion time; TR, repetition time.

same disc level as the MC-related STIR signal. Maximum intensity of the MC-related STIR signal ("Stir") in % points on a scale from normal vertebral body intensity ("Body," 0%) to CSF intensity ("CSF," 100%) was calculated as ((Stir – Body)/(CSF – Body)) × 100.

Statistical analyses

For each endplate L4–S1, we calculated Fleiss' kappa for all observers and Cohen's kappa for each observer pair. Kappa was unweighted for dichotomous variables and linearly weighted for ordinal variables. McNemar's test was applied to compare the prevalence of findings between observers. We computed means of differences between observers with 95% limits of agreement for height and AP extent of findings in % points of vertebral body height and AP extent, and for STIR signal intensity in % points on the scale from normal vertebral body intensity (0%) to CSF intensity (100%). We used sample size weighted means and pooled limits of agreement from all endplates to compute the mean differences between the observers with 95% limits of agreement for all endplates. The 95% limits of agreements represent the limits within which 95% of the differences are expected to occur. We used MedCalc 17.6 (MedCalc Software) to compute means, R 3.5 (R Foundation for Statistical Computing) for kappa and weighted means, and Matlab 9.5 (Mathworks) to derive forest plots.

Only data from L4/L5 and L5/S1 were analyzed, due to <10% prevalence of MCs at higher levels. Kappa is usually not reported for findings with prevalence <10%, as very low prevalence can lead to very low kappa values despite very high actual agreement (46). Kappa was interpreted as: $k \leq 0.20$ = poor; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = good; and 0.81–1.00 = very good agreement beyond chance (47).

Sample size

Assuming a finding has a prevalence of 30%, 85 patients are needed to detect ($\beta = 0.2$, two-sided $\alpha = 0.05$) an unweighted pairwise kappa value of 0.70 as significantly larger than 0.40 (46). We used three observers and 120 patients to further improve the power and increase the size of subgroups. In general, at least 50 individuals are recommended in reliability studies (48).

Results

The observers reported MC-related STIR signal increases regardless of MC type on T1/T2. There were no missing data.

Categorical STIR variables

Overall agreement between the three radiologists (mean Fleiss' kappa) was very good for presence of

Table 2. Criteria for evaluating MCs and related STIR signal increases.

Variables	Description, criteria
MC characteristics evaluated on sagittal T1W and T2W images, blinded to STIR images	
Type	Primary (most extensive) and secondary MC types rated as type I (hypo-intense on T1, hyper-intense on T2), type II (hyperintense on T1, iso- or hyperintense on T2), or type III (hypointense on T1 and T2). Borderline type I vs. type II MCs (near iso-intense on T1) are rated as type II (i.e. type I requires a clearly hypo-intense region on T1)
Height	Largest height of MC measured in mm and rated as <10%, <25%, 25–50%, or >50% of vertebral body height in mm. Both heights are measured along the same line on the same image, excluding the thin low-intensity cortical borders between the bone marrow and the discs.* The <10% category also includes MCs with diameter ≤5 mm
AP extent	Largest AP extent of MC measured in mm and rated as <25%, 25–50%, or >50% of the mid-sagittal AP diameter of the endplate measured in mm
Volume	MC volume subjectively estimated to <10%, <25%, 25–50%, or >50% of total vertebral body marrow volume, taking into account the affected area on all images
STIR signal increase (MC-related), evaluated with T1W/TW2 images available	
Presence	Presence of visible STIR signal increase compared to normal vertebral bone marrow, in relation to MCs seen on T1W/T2W images – or located and shaped as MCs. Rated as no, inside MC, in- and outside MC, or outside MC
Height	Largest height of the region with high STIR signal measured in mm and rated as <10%, <25%, 25–50%, or >50% of vertebral body height in mm. STIR signal height and vertebral body height are both measured along the same line on the same image, excluding the low-intensity cortical borders between the bone marrow and the discs*
AP extent	Largest AP extent of the high STIR signal measured in mm and rated as <25%, 25–50%, or >50% of the mid-sagittal AP diameter of the endplate measured in mm
Volume	Volume of the high STIR signal subjectively rated as <10%, <25%, 25–50%, or >50% of total vertebral body volume, taking into account the affected area on all images
Intensity	Maximum intensity of the high STIR signal, measured in a 25 mm^2 ROI
CSF intensity	STIR signal intensity in the CSF at the level of the vertebral unit with high STIR signal, measured in a 25 mm^2 ROI on the mid-sagittal image, or the next image left or right, avoiding non-CSF structures. If possible, the CSF signal is measured behind the lower half of the cranial vertebra of the vertebral unit (e.g. behind L4 in the L4/L5 unit, if the MC-related STIR signal increase is superior and/or inferior to the L4/L5 disc)
Vertebral body intensity	STIR signal intensity in normal (on STIR, T1 and T2) vertebral body marrow, measured in a 44 mm^2 ROI near the endplate in the central AP third of the opposite normal part (caudal or cranial) of the vertebra with high STIR signal. If this part is not normal, and always when the high STIR signal is in S1, the nearest vertebra above is used for measurement, its caudal part if possible, otherwise its cranial part. The measurement is first considered in the midsagittal image and the next image left or right, before a new location may be considered. The central vertebral vein is not included in the ROI

*In S1 laterally, if the image intended for measuring vertebral body height does not show the S1/S2 interface, the next more medial image is used for this measurement.

AP, anteroposterior; CSF, cerebrospinal fluid; MC, Modic change; ROI, region of interest; STIR, short tau inversion recovery.

MC-related STIR signal increase (0.86), and moderate for its height (0.51), AP extent (0.48), and volume (0.56) (Fig. 2). Kappa values were generally lower at L5/S1 inferior to disc.

The prevalence of STIR signal increase differed maximum 11.6% points between observers (observers A vs. B reported prevalence of 58.3% vs. 46.7% at L5/S1 inferior to disc, $P < 0.001$).

Mean pairwise Cohens' kappa across all categorical STIR variables and levels indicated slightly better agreement between observers A and B (0.62) versus A and C (0.52) and B and C (0.52). Further pairwise STIR results are found in the Appendix (Suppl. Fig. 1, Suppl. Table 2).

Numerical STIR variables

For height of the region with STIR signal increase in % of vertebral body height, the largest mean of differences between observers was 6.9% points and the widest limits of agreement were ±22.3% points, based on data from all levels (Fig. 3). For AP extent of the increased STIR signal in % of AP endplate diameter, the corresponding numbers were 11.2%

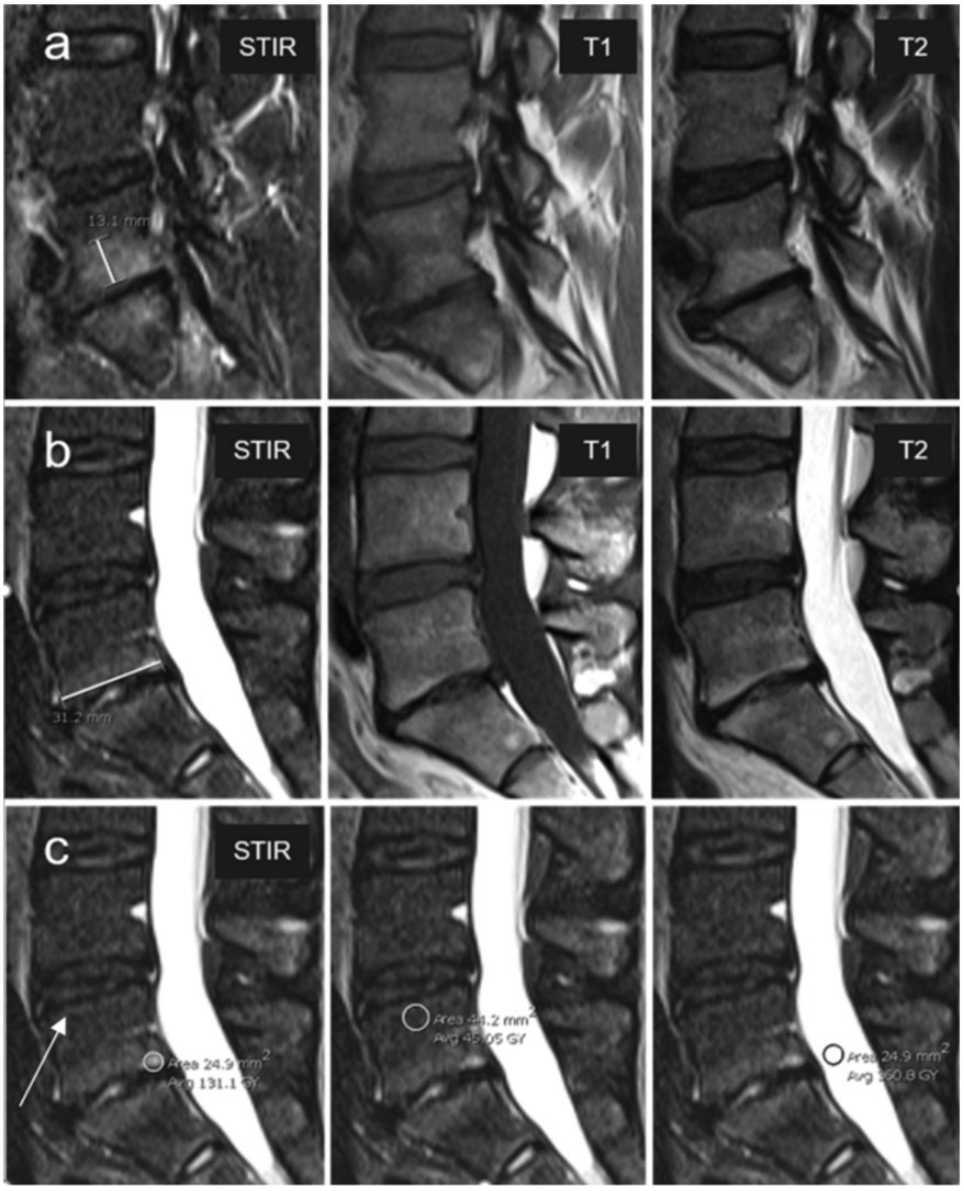

Fig. 1. (a–c) STIR signal increases related to MCs. A 48-year old woman with type II MCs and MC-related STIR signal increases superior and inferior to the L5/S1 disc. The figure shows measurements of (a) height, (b) AP extent, and (c) maximum intensity of the STIR signal with vertebral body and CSF intensities for reference. (c) The circular regions of interest used for measurements are visible with their sizes and gray-scale values, from left to right: maximum intensity (area = 24.9 mm^2, average = 131.1 GY), vertebral body intensity (area = 44.2 mm^2, average = 45.05 GY), and CSF intensity (area = 24.9 mm^2, average = 360.8 GY). Corresponding T1W/T2W fast spin-echo images show type II MCs. Note the diffuse outline of the STIR signal. Note also the thin hyperintense zone on STIR near normal endplates (arrow), which may be mistaken for an AP continuation of an MC-related STIR signal increase at endplates with such increase. AP, anteroposterior; CSF, cerebrospinal fluid; MC, Modic change; STIR, short tau inversion recovery.

points ±34.5% points. For maximum intensity of the STIR signal in % on the scale from normal vertebral body intensity (0%) to CSF intensity (100%), the largest mean of differences and widest limits of agreement were 0.9% and ±7.6% points, based on data from all levels. Results for individual levels are provided in the Appendix (Suppl. Fig. 2).

Reported % points were in the range of 8–100 (mean = 43) for height, 7–100 (mean = 74) for AP extent, and 6–78 (mean = 32) for intensity of STIR signal increases.

MC evaluation on T1/T2

On T1/T2, agreement (mean Fleiss' kappa) was very good for presence of MCs (0.88) and for presence of

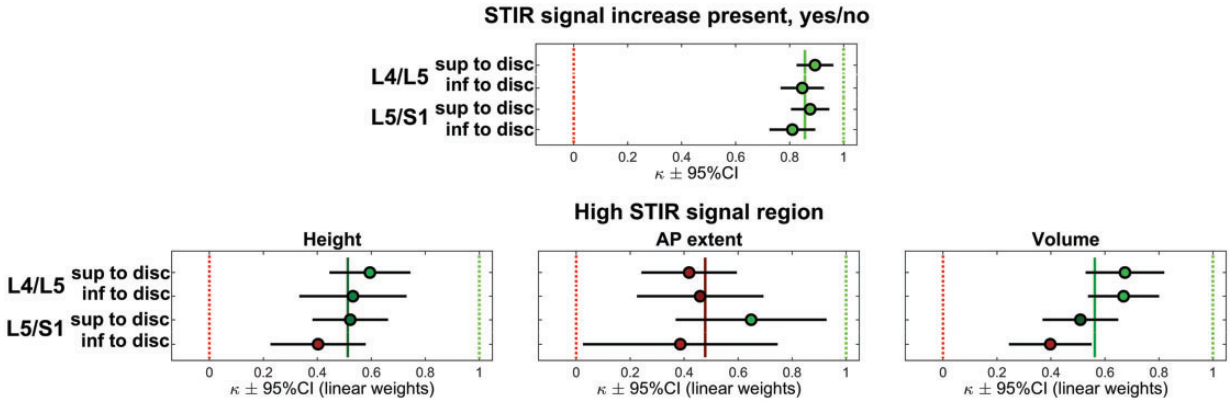

Fig. 2. Categorical STIR variables: forest plot for kappa values with 95% CIs. The figure shows Fleiss' kappa values with 95% CIs for all observers for variables describing MC-related STIR signal increases superior (sup) and inferior (inf) to the L4/L5 and L5/S1 discs. These variables were presence (yes/no), height (four categories), AP extent (three categories), and volume (four categories) of region with high STIR signal. Mean kappa value for agreement between all raters across all four levels L4–S1 is marked with a bold vertical line. This line and circles representing kappa values are green for kappa values >0.50 (the midpoint of the moderate agreement category) and otherwise red. AP, anteroposterior; CI, confidence interval; MC, Modic change; STIR, short tau inversion recovery.

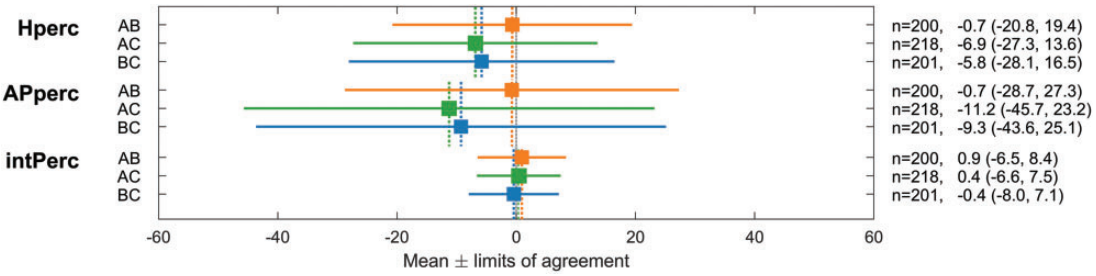

Fig. 3. Numerical STIR variables: forest plot for means of differences and limits of agreement. The figure shows mean of differences with 95% limits of agreement in observer pairs A/B, A/C, and B/C for three numerical variables describing STIR signal increases related to MCs. Each variable was evaluated at four endplates (superior and inferior to the L4/L5 and L5/S1 discs). Means for all endplates are displayed. Values are % points. Hperc denotes height of region with high STIR signal in % of the height of the vertebral body marrow; APperc denotes AP extent of the high STIR signal in % of the mid-sagittal AP diameter of the endplate; intPerc denotes maximum intensity of the STIR signal in % on a scale from normal vertebral body marrow intensity (0%) to CSF intensity (100%). AP, anteroposterior; CSF, cerebrospinal fluid; MC, Modic change; STIR, short tau inversion recovery.

primary or secondary type I MCs (0.81) (Fig. 4). Mean kappa was 0.64 for height, 0.56 for AP extent, and 0.69 for volume of MCs on T1/T2. These values were 0.08–0.13 higher than the corresponding kappa values for dimensions of MC-related STIR signal increases.

The largest difference between observers in prevalence of MCs on T1/T2 was 6.7% points (observers A vs. B reported prevalence 79.2% vs. 72.5% at L5/S1 inferior to disc, $P = 0.021$).

Mean pairwise Cohens' kappa across all categorical T1/T2 variables and levels indicated similar agreement between observers A and B (0.73), A and C (0.71), and B and C (0.73).

The largest mean of differences (and widest limits of agreement) on T1/T2 were for MC height 0.7 (±17.4) % points and for AP extent of MCs 2.6 (±28.8) % points (Fig. 5). These values were smaller than the corresponding values on STIR. On T1/T2, reported % points were in the range of 5–91 (mean = 39) for height and 8–100 (mean = 78) for AP extent.

Further T1/T2 results are detailed in the Appendix (Suppl. Figs. 3 and 4, Suppl. Table 3).

Discussion

To our knowledge, this was the first comprehensive study of the inter-observer reliability for evaluations of STIR signal increases in the vertebral bone marrow. Three radiologists evaluated MC-related high-intensity regions on STIR in 120 patients. Overall inter-observer agreement was very good for the presence of STIR signal increase and moderate for its height, AP extent, and volume. In general, % measured height of the STIR signal differed less

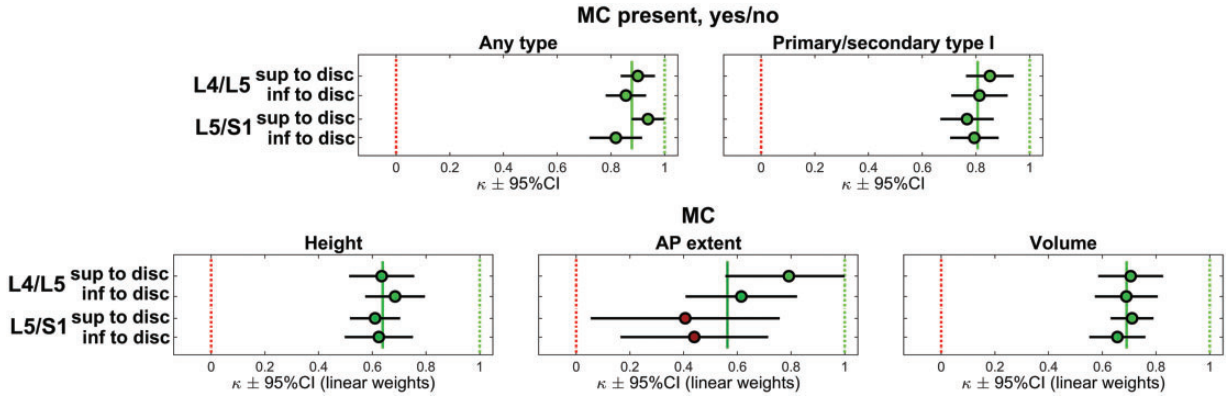

Fig. 4. Categorical MC variables on T1/T2: forest plot for kappa values with 95% CIs. The figure shows Fleiss' kappa values with 95% CIs for all observers for variables describing MCs superior (sup) and inferior (inf) to the L4/L5 and L5/S1 discs on T1W/T2W fast spin-echo images. These variables were presence of any type of MCs (yes/no), presence of primary or secondary type I MCs (yes/no), height (four categories), AP extent (three categories), and volume (four categories) of the MCs. Mean kappa value for agreement between all raters across all four levels L4–S1 is marked with a bold vertical line. This line and circles representing kappa values are green for kappa values >0.50 (the midpoint of the moderate agreement category) and otherwise red. AP, anteroposterior; CI, confidence interval; MC, Modic change; STIR, short tau inversion recovery.

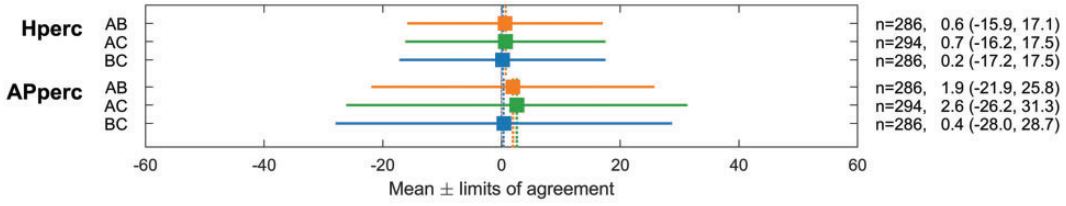

Fig. 5. Numerical MC variables on T1/T2: forest plot for means of differences and limits of agreement. The figure shows mean of differences with 95% limits of agreement in observer pairs A/B, A/C, and B/C for two numerical variables describing MCs on T1W/T2W fast spin-echo images. Each variable was evaluated at four endplates (superior and inferior to the L4/L5 and L5/S1 discs). Means for all endplates are displayed. Values are % points. Hperc denotes height of the MC in % of the height of the vertebral body marrow; APperc means AP extent of the MC in % of the mid-sagittal AP diameter of the endplate. AP, anteroposterior; MC, Modic change.

between observers than its % measured AP extent. For its maximum intensity (on a scale of 0–100%), mean of differences was <1% points and limits of agreement within ±7.6% points.

Further, we have found only one previous study (on 25 patients) of the reliability for MC evaluations on any fat-suppressed fluid-sensitive series (7). In that study, kappa for inter-observer agreement was 0.74 for presence and 0.80 for categorized height of MC-related signal increase. Our corresponding kappa values were 0.86 and 0.51. In patients with possible vertebral fractures (6), kappa for inter-observer agreement on categorized volume of vertebral bone edema on STIR was 0.58; our value was 0.56. It was not reported in these previous studies whether kappa was weighted or categories were combined, so it is not clear whether their kappa values are comparable to ours. In line with our results for intensity measurements, inter-observer agreement was excellent for measurements of signal intensities in vertebral hemangiomas on STIR/fat-saturated T2 images (intra-class correlations of 0.97–0.99) (8).

The very good agreement on presence of STIR signal increases in our study is reassuring for clinical work and research. However, the moderate agreement on the extent of the high STIR signal is not optimal. Moderate inter-observer agreement is common in spine imaging (36,41,49,50), but it implies lowered accuracy when associations with clinical factors are sought (37). More reliable conclusive MRI findings can be based on different observers' separate evaluations followed by their joint conclusion (51). Furthermore, in order to improve agreement between observers, reasons for disagreement should be identified and addressed.

Reasons for disagreement on extent of STIR signal increases may be diffuse outline/gradual lessening of the signals and inhomogeneous bone marrow signal, especially in S1 (where agreement was slightly poorer) (Fig. 2). The AP extent of the MC-related STIR signal often tapers gradually and may blend into a normal

thin hyperintense zone beneath the bony endplate (Fig. 1). This can partly explain larger disagreement for AP extent than for height. The generally larger AP extent than height of the STIR signal is not a likely explanation, as differences in % measured AP extent of the signal were similar for small and large extents (data not shown). On T1/T2, better agreement was achieved for MC extent both in our study and between other experienced observers (36,40,49,50). Therefore, disagreement on extent of MC-related STIR signal increase is probably due to genuine difficulties in interpretation.

Our study also added new information regarding the detection of any area with type I MCs on T1/T2. Previous studies have focused on primary MC types (38–41). We found very good agreement on presence of any (primary or secondary) type I MCs. This may be partly because we rated borderline type I versus type II MCs with near isointense T1 signal (no clear edema) as type II. MCs that are isointense on T1 (and hyperintense on T2) fall outside the original definition of MC types, and it is unclear how they were classified in other studies.

The strengths of this study include the use of three observers (all experienced radiologists), a pilot study, a large sample, standardized MRI protocols and rating criteria, and inclusion of measurements. Multiple observers improved the power and the generalizability of results, and including more patients rather than more than three observers is an effective strategy for maximizing power (46). The data on reliability for T1/T2 evaluations of MCs supported the credibility of our STIR results. We also standardized MC-related STIR signal intensities against normal bone marrow and CSF at the same or a close level, since intensity values varied both between and within patients and depended on craniocaudal and AP localization.

There are also limitations to the study. It was restricted to patients with previously reported MCs and disc herniation. We would however expect similar reliability for STIR evaluations in other patients with low back pain. Intra-observer reliability was not examined; it is often better than the inter-observer reliability (41,43,49,50). A single type of 1.5-T MRI scanner was used, and the results may not be transferrable to images with a different quality or to scanners with a different field strength. Lesion volume was not measured; it was categorized by taking into account (summing up) the visually estimated affected area on all images. Although precise measurements on all images is less feasible, reliability data also for measured volume would have been useful. Finally, all observers had >10 years of experience in musculoskeletal MRI and the reliability for less specialized or less experienced radiologists is still unknown.

We propose the following implications of our results. First, radiologists can evaluate STIR signal increases in the lumbar spine based on criteria used in this study. Second, clinicians and radiologists can expect more reliable evaluation of the height versus the AP extent of a region with MC-related STIR signal increase. Third, when relevant, radiologists can grade the volume of STIR signal increases with reasonable inter-observer reliability, without performing very time-consuming measurements. Fourth, one should still attempt to improve the reliability for size evaluations of STIR signal increases in research, e.g. by joint pre-training, semi-automated lesion contouring (37), continuous volume measurements, and basing conclusive findings on multiple observers' evaluations (51). Fifth, maximum STIR signal intensity relative to normal bone marrow and CSF is an attractive variable in further research, due to its excellent inter-observer reliability. Finally, radiologists can use criteria from this study to improve the evaluation of type I MCs on T1W/T2W fast spin-echo images.

In conclusion, the agreement between experienced radiologists was very good regarding the presence of MC-related STIR signal increase and its maximum intensity, and moderate for its extent and volume. These results provide a basis for validating the relevance of such STIR signal increases for symptoms and treatment results.

Acknowledgments

The authors thank their collaborators in the AIM study group for their contributions.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received the following financial support for the research, authorship, and/or publication of this article: This work was supported by the South East Norway Regional Health Authority (grant no. 2015-090) and the Western Norway Regional Health Authority (grant nos. HV 911891 and HV 911938). They had no role in study design; collection, analysis, and interpretation of data; writing of the report; or the decision to submit the report for publication.

The Regional Committees for Medical Research Ethics in Norway approved this study and the trial it was based on (REC South East, approval no. 2014/158). The trial is registered at ClinicalTrials.gov (identifier NCT02323412).

ORCID iD

Per Martin Kristoffersen https://orcid.org/0000-0001-6471-8861

Supplemental Material

Supplemental material for this article is available online.

References

1. Maksymowych WP, Crowther SM, Dhillon SS, et al. Systematic assessment of inflammation by magnetic resonance imaging in the posterior elements of the spine in ankylosing spondylitis. Arthritis Care Res (Hoboken) 2010;62:4–10.
2. Agten CA, Zubler V, Rosskopf AB, et al. Enthesitis of lumbar spinal ligaments in clinically suspected spondyloarthritis: value of gadolinium-enhanced MR images in comparison to STIR. Skeletal Radiol 2016;45:187–195.
3. Baraliakos X, Hermann KG, Landewe R, et al. Assessment of acute spinal inflammation in patients with ankylosing spondylitis by magnetic resonance imaging: a comparison between contrast enhanced T1 and short tau inversion recovery (STIR) sequences. Ann Rheum Dis 2005;64:1141–1144.
4. Song IH, Hermann KG, Haibel H, et al. Inflammatory and fatty lesions in the spine and sacroiliac joints on whole-body MRI in early axial spondyloarthritis–3-Year data of the ESTHER trial. Semin Arthritis Rheum 2016;45:404–410.
5. Lecouvet FE, Vander Maren N, Collette L, et al. Whole body MRI in spondyloarthritis (SpA): Preliminary results suggest that DWI outperforms STIR for lesion detection. Eur Radiol 2018;28:4163–4173.
6. Diekhoff T, Engelhard N, Fuchs M, et al. Single-source dual-energy computed tomography for the assessment of bone marrow oedema in vertebral compression fractures: a prospective diagnostic accuracy study. Eur Radiol 2019;29:31–39.
7. Finkenstaedt T, Del Grande F, Bolog N, et al. Modic type 1 changes: detection performance of fat-suppressed fluid-sensitive MRI sequences. RoFo 2017;190:152–160.
8. Nabavizadeh SA, Mamourian A, Schmitt JE, et al. Utility of fat-suppressed sequences in differentiation of aggressive vs typical asymptomatic haemangioma of the spine. Br J Radiol 2016;89:20150557.
9. Spirig JM, Sutter R, Gotschi T, et al. Value of standard radiographs, computed tomography, and magnetic resonance imaging of the lumbar spine in detection of intraoperatively confirmed pedicle screw loosening-a prospective clinical trial. Spine J 2019;19:461–468.
10. Modic MT, Ross JS, Masaryk TJ. Imaging of degenerative disease of the cervical spine. Clin Orthop 1989;239:109–120.
11. Modic MT, Steinberg PM, Ross JS, et al. Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology 1988;166:193–199.
12. Modic MT, Ross JS. Lumbar degenerative disk disease. Radiology 2007;245:43–61.
13. Brinjikji W, Diehn FE, Jarvik JG, et al. MRI findings of disc degeneration are more prevalent in adults with low back pain than in asymptomatic controls: a systematic review and meta-analysis. AJNR Am J Neuroradiol 2015;36:2394–2399.
14. Jensen TS, Karppinen J, Sorensen JS, et al. Vertebral endplate signal changes (Modic change): a systematic literature review of prevalence and association with non-specific low back pain. Eur Spine J 2008;17:1407–1422.
15. Zhang YH, Zhao CQ, Jiang LS, et al. Modic changes: a systematic review of the literature. Eur Spine J 2008;17:1289–1299.
16. Herlin C, Kjaer P, Espeland A, et al. Modic changes-Their associations with low back pain and activity limitation: A systematic literature review and meta-analysis. PLoS One 2018;13:e0200677.
17. Hanimoglu H, Cevik S, Yilmaz H, et al. Effects of Modic type 1 changes in the vertebrae on low back pain. World Neurosurg 2019;121:e426–e432.
18. Splendiani A, Bruno F, Marsecano C, et al. Modic I changes size increase from supine to standing MRI correlates with increase in pain intensity in standing position: uncovering the "biomechanical stress" and "active discopathy" theories in low back pain. Eur Spine J 2019;28:983–992.
19. Maatta JH, Karppinen J, Paananen M, et al. Refined phenotyping of Modic changes: imaging biomarkers of prolonged severe low back pain and disability. Medicine 2016;95:e3495.
20. Dudli S, Fields AJ, Samartzis D, et al. Pathobiology of Modic changes. Eur Spine J 2016;25:3723–3734.
21. Beaudreuil J, Dieude P, Poiraudeau S, et al. Disabling chronic low back pain with Modic type 1 MRI signal: acute reduction in pain with intradiscal corticotherapy. Ann Phys Rehabil Med 2012;55:139–147.
22. Buttermann GR. The effect of spinal steroid injections for degenerative disc disease. Spine J 2004;4:495–505.
23. Mefford J, Sairyo K, Sakai T, et al. Modic type I changes of the lumbar spine in golfers. Skeletal Radiol 2011;40:467–473.
24. Cao P, Jiang L, Zhuang C, et al. Intradiscal injection therapy for degenerative chronic discogenic low back pain with end plate Modic changes. Spine J 2011;11:100–106.
25. Albert HB, Sorensen JS, Christensen BS, et al. Antibiotic treatment in patients with chronic low back pain and vertebral bone edema (Modic type 1 changes): a double-blind randomized clinical controlled trial of efficacy. Eur Spine J 2013;22:697–707.
26. Cai G, Laslett LL, Aitken D, et al. Effect of zoledronic acid and denosumab in patients with low back pain and Modic change: a proof-of-principle trial. J Bone Miner Res 2018;33:773–782.
27. Fayad F, Lefevre-Colau MM, Rannou F, et al. Relation of inflammatory modic changes to intradiscal steroid

injection outcome in chronic low back pain. Eur Spine J 2007;16:925–931.
28. Koivisto K, Jarvinen J, Karppinen J, et al. The effect of zoledronic acid on type and volume of Modic changes among patients with low back pain. BMC Musculoskelet Disord 2017;18:274.
29. Koivisto K, Kyllonen E, Haapea M, et al. Efficacy of zoledronic acid for chronic low back pain associated with Modic changes in magnetic resonance imaging. BMC Musculoskelet Disord 2014;15:64.
30. Al-Falahi MA, Salal MH, Abdul-Wahab DM. Antibiotic Treatment in Patients with Chronic Low Back Pain and Vertebral Bone Edema (Modic Type I Changes): A Randomized Clinical Controlled Trial of Efficacy. Iraqi Postgraduate Medical Journal 2014;13:390–398.
31. Palazzo C, Ferrari M, Lefevre-Colau M-M, et al. Lack of effectiveness of antibiotics in chronic low back pain with Modic 1 changes. Joint Bone Spine 2017;84:507–508.
32. Storheim K, Espeland A, Grovle L, et al. Antibiotic treatment In patients with chronic low back pain and Modic changes (the AIM study): study protocol for a randomised controlled trial. Trials 2017;18:596.
33. Wilkens P, Storheim K, Scheel I, et al. No effect of 6-month intake of glucosamine sulfate on Modic changes or high intensity zones in the lumbar spine: sub-group analysis of a randomized controlled trial. J Negat Results Biomed 2012;11:13.
34. Braten LCH, Rolfsen MP, Espeland A, et al. Efficacy of antibiotic treatment in patients with chronic low back pain and Modic changes (the AIM study): double blind, randomised, placebo controlled, multicentre trial. BMJ 2019;367:l5654.
35. Feinstein AR. An additional basic science for clinical medicine: IV. The development of clinimetrics. Ann Intern Med 1983;99:843–848.
36. Jarvik JG, Deyo RA. Moderate versus mediocre: the reliability of spine MR data interpretations. Radiology 2009;250:15–17.
37. Fields AJ, Battié MC, Herzog RJ, et al. Measuring and reporting of vertebral endplate bone marrow lesions as seen on MRI (Modic changes): recommendations from the ISSLS Degenerative Spinal Phenotypes Group. Eur Spine J 2019;28:2266–2274
38. Jensen TS, Sorensen JS, Kjaer P. Intra- and interobserver reproducibility of vertebral endplate signal (modic) changes in the lumbar spine: the Nordic Modic Consensus Group classification. Acta Radiol 2007;48:748–754.
39. Jones A, Clarke A, Freeman BJ, et al. The Modic classification: inter- and intraobserver error in clinical practice. Spine 2005;30:1867–1869.
40. Peterson CK, Gatterman B, Carter JC, et al. Inter- and intraexaminer reliability in identifying and classifying degenerative marrow (Modic) changes on lumbar spine magnetic resonance scans. J Manipulative Physiol Ther 2007;30:85–90.
41. Arana E, Royuela A, Kovacs FM, et al. Lumbar spine: agreement in the interpretation of 1.5-T MR images by using the Nordic Modic Consensus Group classification form. Radiology 2010;254:809–817.
42. Tibiletti M, Ciavarro C, Bari V, et al. Semi-quantitative evaluation of signal intensity and contrast-enhancement in Modic changes. Eur Radiol Exp 2017;1:5.
43. Wang Y, Videman T, Niemelainen R, et al. Quantitative measures of modic changes in lumbar spine magnetic resonance imaging: intra- and inter-rater reliability. Spine 2011;36:1236–1243.
44. Kottner J, Audige L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 2011;64:96–106.
45. Fardon DF, Williams AL, Dohring EJ, et al. Lumbar disc nomenclature: version 2.0: Recommendations of the combined task forces of the North American Spine Society, the American Society of Spine Radiology and the American Society of Neuroradiology. Spine J 2014;14:2525–2545.
46. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005;85:257–268.
47. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
48. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42.
49. Berg L, Neckelmann G, Gjertsen O, et al. Reliability of MRI findings in candidates for lumbar disc prosthesis. Neuroradiology 2012;54:699–707.
50. Carrino JA, Lurie JD, Tosteson AN, et al. Lumbar spine: reliability of MR imaging findings. Radiology 2009;250:161–170.
51. Espeland A, Vetti N, Krakenes J. Are two readers more reliable than one? A study of upper neck ligament scoring on magnetic resonance images. BMC Med Imaging 2013;13:4.