British Society for Rheumatology

RHEUMATOLOGY

OXFORD

Clinical science

Self-reported diagnostic confidence predicts diagnostic accuracy in axial spondyloarthritis imaging

Sevtap Tugce Ulas [1,2,*], Felix Radny[1], Katharina Ziegeler [1], Iris Eshed [3], Juliane Greese[1], Dominik Deppe[1], Carsten Stelbrink[1], Robert Biesen [4], Hildrun Haibel [5], Valeria Rios Rodriguez [5], Judith Rademacher[2,5], Mikhail Protopopov [5], Fabian Proft [5], Denis Poddubnyy [5], Torsten Diekhoff [1,2]

[1]Department of Radiology, Charité – Universitätsmedizin Berlin, Campus Mitte, Humboldt-Universität zu Berlin, Freie Universität Berlin, Berlin, Germany
[2]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany
[3]Department of Diagnostic Imaging, Sheba Medical Center, affiliated with the Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel
[4]Department of Rheumatology, Charité – Universitätsmedizin Berlin, Campus Mitte, Humboldt-Universität zu Berlin, Freie Universität Berlin, Berlin, Germany
[5]Department of Rheumatology, Charité – Universitätsmedizin Berlin, Campus Benjamin Franklin, Humboldt-Universität zu Berlin, Freie Universität Berlin, Berlin, Germany
*Correspondence to: Sevtap Tugce Ulas, Department of Radiology (CCM), Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany. E-mail: sevtap-tugce.ulas@charite.de

Abstract

Objectives: Reporting diagnostic confidence (DC) in axial spondyloarthritis (axSpA) imaging is recommended by the ASAS guidelines. Our aim was to investigate whether self-reported DC predicts diagnostic accuracy in axSpA imaging using X-ray (XR), computed tomography (CT) and magnetic resonance imaging (MRI).

Methods: We performed a *post hoc* analysis including 163 patients with low back pain (89 axSpA and 56 non-axSpA). Nine blinded readers with different experience levels [inexperienced ($<$1 year), semi-experienced (3–8 years) and experienced ($>$12 years)] scored the sacroiliac joint images for compatibility with axSpA. DC was reported on a scale from 1 (not sure) to 10 (very sure). Mean DC scores and standard deviations were calculated for correct and incorrect responses using XR, CT, MRI, XR+MRI and CT+MRI. Differences in DC were assessed using the Mann–Whitney U test.

Results: DC scores were higher for correct axSpA diagnoses and differed significantly between correct and incorrect responses for all modalities ($P < 0.001$), with a mean DC of 7.1 ± 2.1 and 6.3 ± 2.1 for XR, 8.3 ± 1.8 and 6.7 ± 2.0 for CT, 8.1 ± 1.9 and 6.2 ± 1.9 for MRI, 8.2 ± 1.8 and 6.7 ± 1.8 for XR+MRI and 8.4 ± 1.8 and 6.8 ± 1.8 for CT+MRI, respectively. This was also the case when looking at the results by experience group, except for XR in the inexperienced group.

Conclusion: Providing self-reported DC in radiological reports is useful information to predict diagnostic reliability in axSpA imaging.

Keywords: magnetic resonance imaging, computed tomography, axial spondyloarthritis.

Rheumatology key messages

- Reporting the diagnostic confidence (DC) in the diagnosis of axSpA in the radiological report is recommended by the current ASAS guidelines.
- A self-reported DC score can be considered a predictive parameter for the readers' diagnostic performance in the evaluation of patients with axSpA.
- Based on our findings, we hypothesize that the addition of DC in radiology reports is valuable information to support clinical decision-making.

Introduction

Axial spondyloarthritis (axSpA) is a common chronic, immune-mediated disease predominantly affecting the sacroiliac joints

(SIJs) and the spine [1, 2] and mainly starting in young adults, who typically suffer from low back pain related to inflammation, which might lead to loss of function as the disease

progresses and structural damage occurs [3–5]. Imaging, particularly magnetic resonance imaging (MRI), is of great importance in the diagnosis of axSpA [6] and has a great impact on the clinical diagnostic confidence (DC) of rheumatologists due to the lack of conclusive laboratory parameters and the fact that the axial skeleton is less easily accessible to clinical examination [7–9]. However, it remains a challenge to distinguish between mechanical, degenerative and non-specific causes of low back pain and axSpA either clinically [9, 10] or with imaging [11]. The accurate use of imaging in both initial diagnosis and treatment monitoring is therefore of particular importance for further patient management [6, 12]. Adequate detection of axSpA-compatible changes in the SIJs and spine is essential in clinical practice and can thus contribute to a reliable diagnosis, especially when the DC of the rheumatologists is low due to non-specific clinical and laboratory findings.

Several imaging findings are specific for axSpA. Structural lesions, such as erosions and ankyloses, and acute inflammatory changes, such as osteitis and enthesitis, are most helpful in making the diagnosis, although ankylosis is more suggestive of an advanced stage of the disease. Each imaging modality has advantages and disadvantages in the evaluation of the individual imaging features. X-ray (XR), MRI and computed tomography (CT) are the most important modalities used to detect axSpA [13]. XR of the SIJs and spine is the first-line imaging test to detect structural lesions but has very low sensitivity and specificity [6, 14]. MRI, on the other hand, is crucial for the detection of inflammation, while it may overestimate structural lesions [15, 16]. Moreover, relying on the demonstration of osteitis alone for diagnosis has been criticized in recent years and may also lead to overdiagnosis of axSpA, as SIJ osteitis may be caused by several other conditions [11]. CT, with its high resolution even when performed with low-dose technique, is currently used in cases where either XR or MRI is inconclusive, allowing detection of structural lesions with higher accuracy [14, 15]. However, standard CT is insensitive to active bone and soft tissue inflammation. Given that each imaging technique has its inherent strengths and weaknesses, it is crucial to tailor the appropriate modality to the specific clinical question. The performance and accuracy of radiologists in interpreting imaging findings is an additional important factor for a correct diagnosis and further patient management. Indeed, reader accuracy has been shown to be directly related to the reader's experience [8]. In this context, presence of all changes associated with axSpA may lead to a different level of DC of the radiologist than if, for example, only isolated changes of the SIJs are present.

Published data show that the reading performance is directly related to reader experience [8]. To date, confidence in assessing axSpA-related changes in imaging modalities has not played a relevant part in the interpretation of radiological findings and reports. Nevertheless, self-reported DC can be an important parameter in the interpretation of a radiology report in the clinical context. For this reason, the Assessment of Spondyloarthritis International Society (ASAS) working group has recommended to include the self-reported DC in radiological reports in its current recommendation [17].

The aim of this study was to systematically investigate the usefulness of self-reported DC for assessing reader performance in axSpA imaging using XR, MRI and CT.

Materials and methods

Subjects

The study was designed as a *post hoc* analysis of two prospective study populations of patients with chronic low back pain and suspected or known axSpA: the Sacroiliac Joint Magnetic Resonance Imaging and Computed Tomography (SIMACT) study [18], and the Virtual Non-Calcium—Susceptibility Weighted Imaging (VNC-SWI) study [19]. For the present analysis, patients were divided into two groups: the axSpA group (based on the diagnosis by expert rheumatologists of the local rheumatology department) and the control group of patients with non-axSpA (i.e. patients with degenerative or mechanical SIJ changes and patients with non-specific low back pain). Subjects with missing or incomplete imaging and clinical data were excluded from analysis.

Ethical approval and data availability

Written informed consent was obtained from all patients before enrolment. The institutional ethics review board of Charité Universitätsmedizin Berlin approved all investigations prior to respective study commencement (EA1/0886/16; EA1/073/10). Patients were not involved in study planning.

Readers and scoring system

For all patients included, XR, MRI and CT datasets of the SIJs were available and were separately anonymized prior to scoring. This resulted in five separate datasets per patient with different anonymization: XR, CT, MRI alone, and XR+MRI and CT+MRI in combination. XR, MRI [oblique-coronal T1-weighted and short-tau inversion recovery (STIR) sequences] and CT were scored by nine readers with different years of experience in musculoskeletal imaging divided into three reader groups (RG): the inexperienced RG (RG1) consisted of medical research students with 0–1 year of experience (C.S., F.R. and D.D.), the semi-experienced RG (RG2) consisted of physicians (radiologists) with 3–8 years of experience (S.T.U., J.G. and K.Z.) and the experienced RG (RG3) consisted of senior physicians (two radiologists and one rheumatologist) with 12–17 years of experience in musculoskeletal imaging (T.D., I.E. and D.P.). Each patient's final diagnosis (axSpA or other) determined by the experienced rheumatologist prior to the scoring based on clinical and imaging findings served as the standard of reference. The readers were blinded to clinical data and the image datasets and results of imaging that was not subject to the current scoring. Readers scored the images for the presence *vs* absence of axSpA in two steps: in step 1, readers were asked to decide whether the image was 'normal 'or 'abnormal', and to report the level of DC for their decision [on a scale from 1 (not sure) to 10 (very sure)]. For each image classified as 'abnormal', readers were then asked, in step 2, to provide a diagnosis, either 'axSpA' or 'other', and to again rate their DC (same scale) in choosing the diagnosis (see detailed scoring system in Fig. 1). The scoring was performed separately for each dataset (XR, CT and MRI, XR+MRI and CT+MRI, respectively).

Statistical analysis

Statistical analysis was performed using GraphPad Prism (Version 9.2.0. for MacOS, GraphPad Software, La Jolla, CA, USA). Mean DC scores and standard deviation (SD) were calculated for the correct and incorrect diagnosis of 'axSpA' for each reader separately for XR, CT, MRI, XR+MRI and
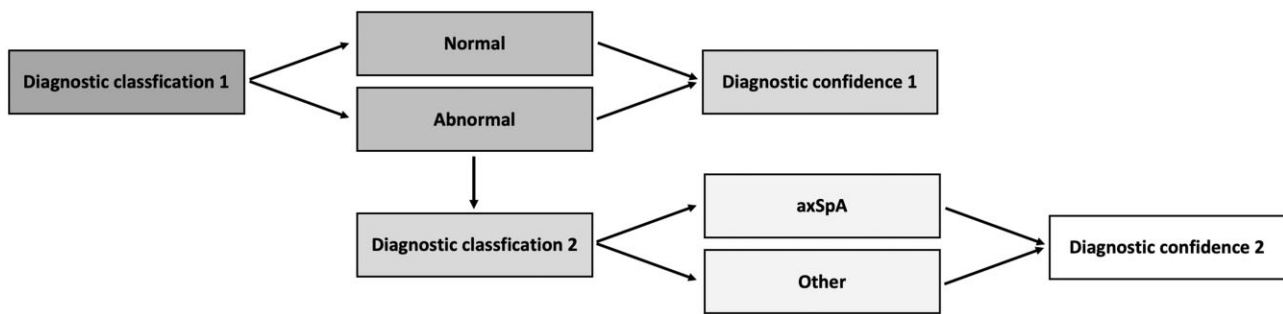
Figure 1. Scoring system and procedure. In step 1, readers were asked to decide whether the images were 'normal' or 'abnormal', indicating their diagnostic confidence (diagnostic confidence 1). For an image classified as 'abnormal', readers were then asked, in step 2, to decide whether the diagnosis was 'axSpA' or 'other', again indicating their diagnostic confidence (diagnostic confidence 2)

CT+MRI. Due to a skewed distribution of DC scores, group differences were assessed using the two-tailed Mann–Whitney *U* test, separately for the three reader groups. Receiver operating characteristics (ROC) analysis was conducted to assess how and whether the DC score can predict the correct diagnosis for axSpA or non-axSpA separately for each modality and combination. To investigate the diagnostic accuracy at various DC levels, we grouped the DC into three levels (level of confidence: LoC), which consist of: low-LoC for DC <4, medium-LoC for DC between 4 and 6, and high-LoC for DC >7. We calculated the percentage of correct and incorrect diagnoses for all readers collectively and for each specific respective reader group separately. Fleiss' kappas (κ) were calculated to determine inter-rater reliability and were interpreted according to Landis and Koch [20]. A *P*-value smaller than 0.05 was considered statistically significant.

Results

Subjects

Overall, 182 patients were evaluated for enrolment, and a total of 163 patients [82 women; mean age of 38 (SD 10.6), 19–62 years] were included in the study after applying the exclusion criteria. Of these, 89 patients were diagnosed with axSpA, 56 with degenerative or mechanical SIJ disease and 18 patients with non-specific low back pain, see Fig. 2. Mean duration of back pain was 6.7 years (SD 7.5). Patients with axSpA had a mean Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) of 4.6 (SD 1.8).

Image reading

Overall, readers' DC differed significantly between correct and incorrect diagnoses in all modalities, with a mean DC score for correct and incorrect diagnoses of 7.1 ± 2.1 and 6.3 ± 2.1 for XR ($P < 0.001$), 8.3 ± 1.8 and 6.7 ± 2.0 for CT ($P < 0.001$), 8.1 ± 1.9 and 6.2 ± 1.9 for MRI ($P < 0.001$), 8.2 ± 1.8 and 6.7 ± 1.8 for XR+MRI ($P < 0.001$) and 8.4 ± 1.8 and 6.8 ± 1.8 for CT+MRI ($P < 0.001$).

The distribution of DC scores between correct and incorrect diagnoses in step 2 by reader group is shown in Fig. 3. In all RGs and modalities, DC scores were higher for correct diagnoses of 'axSpA' while lower DC scores were obtained with incorrect diagnoses of 'axSpA'. DC scores differed significantly for correct and incorrect responses in RG2 and RG3 for all modalities (see Fig. 3). In the inexperienced group (RG1), there was no significant difference in DC between correct and incorrect responses for XR. In addition, the

experienced readers (RG3) showed higher DC scores overall for correct diagnoses of 'axSpA'. The scoring results for step 1 ('normal' *vs* 'abnormal') are presented in Supplementary Figs S1 and S2 (available at *Rheumatology* online).

ROC analysis revealed DC to have an important predictive power for reader performance, except for XR in the inexperienced group (Fig. 4). Furthermore, the addition of XR to MRI (XR+MRI) did not improve the diagnostic confidence in the diagnosis of 'axSpA' in all reader groups as demonstrated by lower area under the curve (AUC) values, see also Fig. 4. The results of the separate analysis of diagnostic accuracy for the specific LoC are summarized in Supplementary Figs S3 and S4 (available at *Rheumatology* online). The agreement of the DC between the readers was highest in the experienced group with moderate to almost perfect agreement, followed by the semi-experienced group with moderate agreement. The inexperienced group exhibited only slight to fair agreement. The highest inter-rater reliability was demonstrated for the experienced readers with CT+MRI ($\kappa = 0.88$), see also Fig. 3 and Supplementary Fig. S1, available at *Rheumatology* online.

Discussion

To the best of our knowledge, this is the first analysis to investigate the relationship between self-reported diagnostic confidence and reader performance in axSpA imaging. We found significant differences in DC between correct and incorrect diagnostic decisions in axSpA made by nine readers with different levels of experience in musculoskeletal imaging. Our study results suggest that using self-reported DC scores to predict diagnostic accuracy of CT and MRI findings is feasible, whereas in XR reading, predicting reader performance based on their self-reported DC remains limited. These results suggest that, in SIJ XR, it is much more difficult for readers to estimate their own performance, especially for inexperienced readers. Furthermore, we show that more experienced readers are better at estimating their own DC and, thus, at cautioning against false conclusions.

Self-reported DC is a promising tool to quantify reader confidence in diagnostic imaging. In the current ASAS working group guidelines for reporting imaging findings in axSpA, the authors recommend to include the level of DC in the radiological report [21]. Our results suggest that self-reported DC has a significant impact on the prediction of performance and can emphasize the confidence. Reporting diagnostic certainty is therefore of crucial importance for further interpretation of
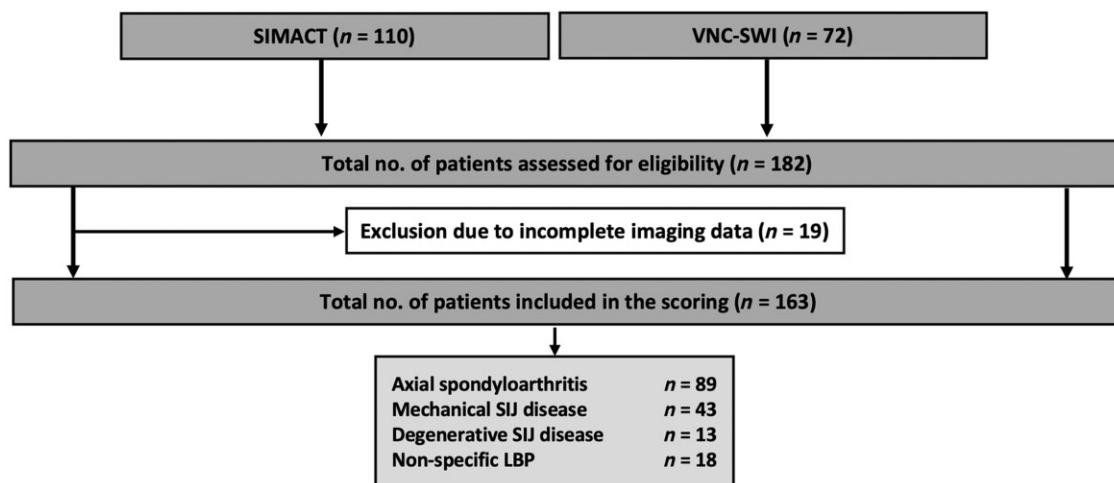
Figure 2. Flowchart of study inclusion. CT: computed tomography; LBP: low back pain; MRI: magnetic resonance imaging; RG: reader group (based on years of experience in musculoskeletal imaging); RG1: inexperienced reader group; RG2: semi-experienced reader group; RG3: experienced reader group; SIMACT: Sacroiliac Joint Magnetic Resonance Imaging and Computed Tomography Study; VNC-SWI: Virtual Non-Calcium—Susceptibility Weighted Imaging study; XR: X-ray
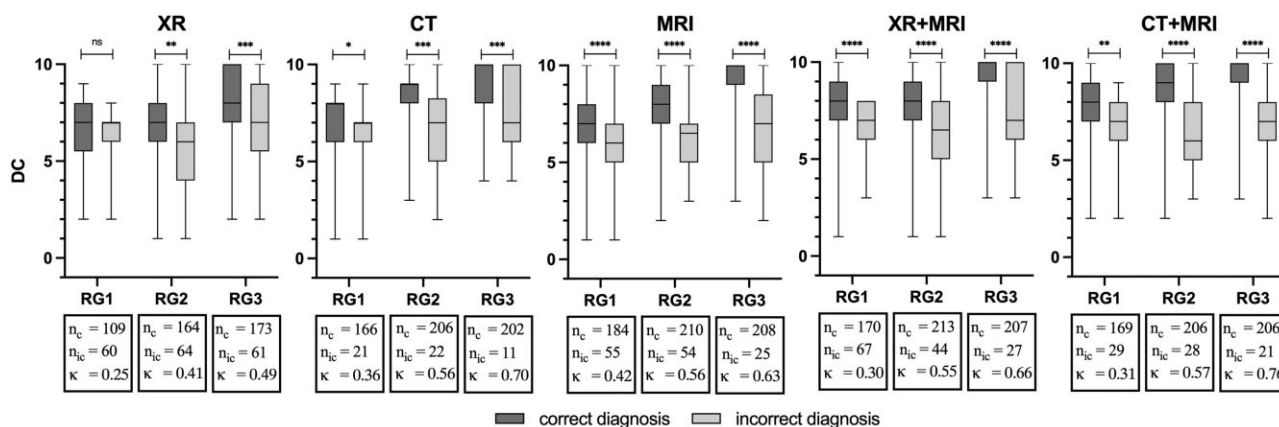

Figure 3. Boxplots of the step 2 scoring results displaying differences in diagnostic confidence between correct and incorrect diagnoses. CT: computed tomography; DC: diagnostic confidence; κ: Fleiss kappa; MRI: magnetic resonance imaging; n_c: number of correct diagnoses; n_{ic}: number of incorrect diagnoses; RG: reader group (based on their years of experience in musculoskeletal imaging); RG1: inexperienced reader group; RG2: semi-experienced reader group; RG3: experienced reader group; XR: X-ray. The symbols (ns: $P > 0.05$, $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$, $****P \leq 0.0001$) represent the results of the conducted Mann–Whitney U test. Except for RG1 in XR, self-reported DC scores were significantly different between correct and incorrect axSpA diagnoses. In addition, the figure also shows that RG3 tended to be more confident in their diagnoses than the two less experienced RGs

the findings. In the clinic, clinical conferences serve to openly discuss the certainty of imaging findings. However, such discussions are rare for physicians in private practice, so that including self-reported DC in radiology reports is of great importance and can prevent misinterpretation of these findings by the rheumatologists.

Readers can estimate and report their performance in various ways. In our study, we chose a more granular categorization of DC levels using a scale from 1 to 10. This fine-tuned score has the advantage of allowing identification of possible cut-off values of DC that might be associated with certain probabilities in terms of diagnostic performance. Indeed, a DC score of 7 or higher turned out to be significantly associated with accurate decision making. Thus, clinicians can derive value from this information by assessing the potential accuracy of the imaging-based diagnosis and incorporating it into their clinical diagnostic decision-making process. However, the range of DC scores was relatively broad within the respective reader groups, with overlaps, indicating that a correct diagnosis could also be associated with a low DC

score and vice versa. DC scores varied among the different reader groups, which should also be considered when weighing the diagnostic report. As there is no easy way to standardize the rater's assessment for DC and it demonstrated a certain dependence on reader's experience and individual preferences, our presented cut-off values should be interpreted with caution. Alternatively, a different DC scaling is also conceivable for clinical practice, in which, for example, a 3-level distinction is made between high, moderate and low DC. This can further increase feasibility in clinical practice and provide better comparability between readers. Furthermore, it needs to be discussed whether it is also useful to report the radiologist's level of experience in the radiological report, which in turn has an impact on DC. However, indicating the level of experience might be more challenging to objectify and, therefore, could be a subject of controversy. On the other hand, the confidence scale is entirely subjective and easy to define. At the same time, the results of this study also indicate, as expected, that training of radiologists and rheumatologists in the assessment of imaging findings is of particular importance
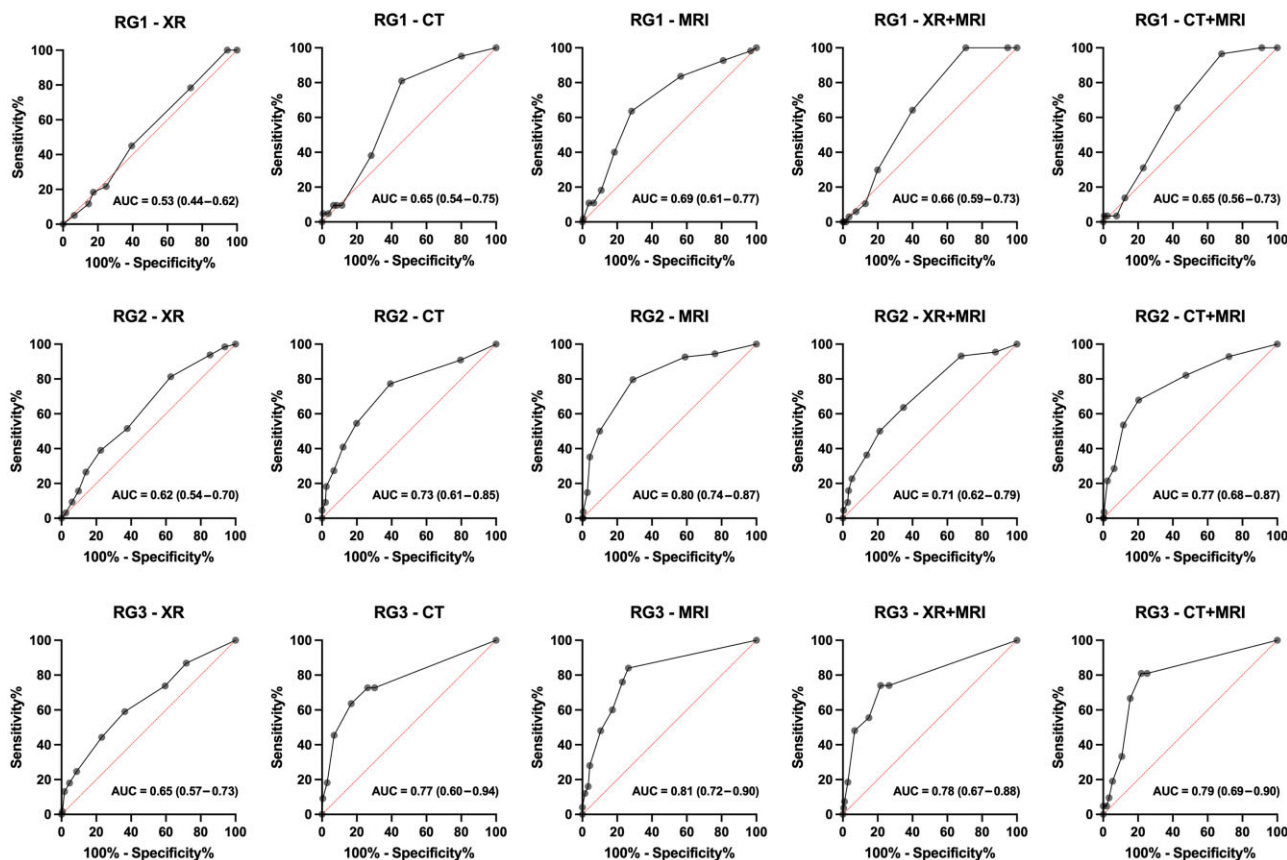
Figure 4. ROC analysis of diagnostic confidence in relation to imaging modality and readers' level of experience (step 2). AUC: area under the ROC curve (95% confidence interval); CT: computed tomography; MRI: magnetic resonance imaging; RG: reader group (based on years of experience in musculoskeletal imaging); RG1: inexperienced reader group; RG2: semi-experienced reader group; RG3: experienced reader group; XR: X-ray. ROC analysis shows that diagnostic confidence has an important predictive power for readers' performance, except for XR in the inexperienced reader group

to improve their own DC, allowing more reliable prediction of performance in axSpA imaging. Awareness of the technical limitations of the used modalities is also a potential contributor to the significant differences in DC between correct and incorrect axSpA diagnoses. These might be better appreciated by the experienced readers, which is then reflected in the reported DC, whereas a higher DC value does not always indicate higher diagnostic accuracy, especially for less experienced readers. A further hypothesis is that DC could reflect the degree of complexity of the case as regards the interpretation of imaging. In cases that are more difficult to interpret, the DC will be lower, and thus there will be the potential of greater discrepancy with the final diagnosis.

Our findings suggest that, in addition to being less sensitive to erosive changes, readers of XR images are less self-aware of their diagnostic confidence and performance and will not be equally able to predict their accuracy as compared with cross-sectional imaging. We believe that this is due to the inherent technical disadvantages of projection radiography as superposition of bowel gas or osteophytes can lead to misinterpretation and false-positive results. Readers are aware of this fact, resulting in overall lower DC in XR and, subsequently, in less accurate prediction of their own diagnostic decision. This adds a further perspective to the current critical discussion about the use of XR in the diagnosis of axSpA [14, 22].

In addition, interestingly, and contrary to expectations, the combination of imaging modalities (MRI+CT and MRI+XR)

did not result in significantly higher diagnostic confidence than MRI alone. A possible explanation may lie in the different degree of difficulty in interpreting these cases, which may represent a confounding factor.

This study was specifically planned to investigate the predictive power of DC for readers' performance. We did this with a total nine readers with different levels of experience in musculoskeletal imaging and investigated DC scores separately for different modalities and combinations to identify differences related to experience and type of imaging modality. Our study has some limitations. Only images of SIJs were included in the scoring and the situation in spine imaging remains to be investigated. Intra-rater reliability was not assessed. Furthermore, readers were blinded to clinical information; access to this information could affect both DC and diagnostic accuracy. The imaging datasets used in this analysis were also used in the routine diagnostic process. This approach does carry the risk of bias from circular reasoning, as some of the radiologists included in the reading may have also been involved in the original reporting in the clinical setting. However, the reading results from this study were not part of the clinical diagnostic process, and given the lack of a widely accepted independent reference standard, this approach is common in axSpA studies. Nonetheless, we mitigated potential recall bias by including a large number of patients and implementing a meticulous anonymization process.

In conclusion, DC allows prediction of readers' performance in axSpA imaging. This is especially true for the

interpretation of cross-sectional imaging modalities such as CT and MRI, which showed significantly higher DC scores for correct diagnoses across all reader experience groups. Therefore, we encourage the inclusion of self-reported DC in the interpretation of imaging findings in axSpA patients into radiology reports.

Supplementary material

Supplementary material is available at *Rheumatology* online.

Data availability

All data and materials presented in this study are available on request from the corresponding author.

References

1. Navarro-Compan V, Sepriano A, El-Zorkany B, van der Heijde D. Axial spondyloarthritis. Ann Rheum Dis 2021;80:1511–21.
2. Sieper J, Poddubnyy D. Axial spondyloarthritis. Lancet 2017;390: 73–84.
3. Sieper J, Rudwaleit M, Baraliakos X *et al.* The Assessment of SpondyloArthritis international Society (ASAS) handbook: a guide to assess spondyloarthritis. Ann Rheum Dis 2009;68(Suppl 2): ii1–44.
4. Sieper J, Rudwaleit M, Khan MA, Braun J. Concepts and epidemiology of spondyloarthritis. Best Pract Res Clin Rheumatol 2006; 20:401–17.
5. Braun J, Bollow M, Remlinger G *et al.* Prevalence of spondylarthropathies in HLA-B27 positive and negative blood donors. Arthritis Rheum 1998;41:58–67.
6. Maksymowych WP, Lambert RG, Ostergaard M *et al.* MRI lesions in the sacroiliac joints of patients with spondyloarthritis: an update of definitions and validation by the ASAS MRI working group. Ann Rheum Dis 2019;78:1550–8.
7. Molto A, Paternotte S, Comet D *et al.* Performances of the Assessment of SpondyloArthritis International Society axial spondyloarthritis criteria for diagnostic and classification purposes in patients visiting a rheumatologist because of chronic back pain: results from a multicenter, cross-sectional study. Arthritis Care Res (Hoboken) 2013;65:1472–81.
8. Poddubnyy D, Sieper J. Diagnostic delay in axial spondyloarthritis—a past or current problem? Curr Opin Rheumatol 2021;33: 307–12.
9. Rudwaleit M, van der Heijde D, Khan MA, Braun J, Sieper J. How to diagnose axial spondyloarthritis early. Ann Rheum Dis 2004;63: 535–43.
10. van der Heijde D, Rudwaleit M, Landewe RB, Sieper J. Justification for including MRI as a tool in the diagnosis of axial SpA. Nat Rev Rheumatol 2010;6:670–2.
11. Diekhoff T, Lambert R, Hermann KG. MRI in axial spondyloarthritis: understanding an 'ASAS-positive MRI' and the ASAS classification criteria. Skeletal Radiol 2022;51:1721–30.
12. Baraliakos X, Ostergaard M, Lambert RG *et al.* MRI lesions of the spine in patients with axial spondyloarthritis: an update of lesion definitions and validation by the ASAS MRI working group. Ann Rheum Dis 2022;81:1243–51.
13. Baraliakos X. Imaging in axial spondyloarthritis. Isr Med Assoc J 2017;19:712–8.
14. Diekhoff T, Eshed I, Radny F *et al.* Choose wisely: imaging for diagnosis of axial spondyloarthritis. Ann Rheum Dis 2022;81: 237–42.
15. Lambert RGW, Hermann KGA, Diekhoff T. Low-dose computed tomography for axial spondyloarthritis: update on use and limitations. Curr Opin Rheumatol 2021;33:326–32.
16. Poddubnyy D, Diekhoff T, Baraliakos X, Hermann KGA, Sieper J. Diagnostic evaluation of the sacroiliac joints for axial spondyloarthritis: should MRI replace radiography? Ann Rheum Dis 2022;81: 1486–90.
17. Diekhoff T, Eshed I, Giraudo C *et al.* OP0150 ASAS recommendations for requesting and reporting imaging examinations in patients with suspected axial spondyloarthritis. Ann Rheum Dis 2022;81:97.
18. Diekhoff T, Greese J, Sieper J *et al.* Improved detection of erosions in the sacroiliac joints on MRI with volumetric interpolated breath-hold examination (VIBE): results from the SIMACT study. Ann Rheum Dis 2018;77:1585–9.
19. Deppe D, Hermann KG, Proft F *et al.* CT-like images of the sacroiliac joint generated from MRI using susceptibility-weighted imaging (SWI) in patients with axial spondyloarthritis. RMD Open 2021;7: e001656.
20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
21. Maksymowych WP, Lambert RG, Baraliakos X *et al.* Data-driven definitions for active and structural MRI lesions in the sacroiliac joint in spondyloarthritis and their predictive utility. Rheumatology (Oxford) 2021;60:4778–89.
22. Eshed I, Diekhoff T, Hermann KGA. Is it time to move on from pelvic radiography as the first-line imaging modality for suspected sacroiliitis? Curr Opin Rheumatol 2023;35:219–25.