1 **Title**

2 The Norwich Osteoarthritis of the Ankle MRI Score (NOAMS); a reliability study.

3

Keywords

- Osteoarthritis
- Ankle
- Magnetic Resonance Imaging
- Reliability

Abstract

Aim

The aim of this study was to define, and test the inter and intra-rater reliability, of a grading system for staging OA of the ankle with MRI (NOAMS).

Materials and Methods

The MR features to be included in the score were defined by a multidisciplinary expert panel through a Delphi process. An anonymised randomised dataset of 50 MR studies was created from patients with concurrent plain radiographs to include 10 ankles of each of the Kellgren-Lawrence grades 0 to 4. Two experienced musculoskeletal radiologists and two trainees scored each ankle MR twice independently and blinded to the plain radiographs.

Results

The inter-rater kappa coefficient of agreement for cartilage disease was 0.88 (95% CI: 0.85, 0.91) for experienced raters and 0.71 (95% CI: 0.67, 0.76) for trainees. Inter-rater agreement for subchondral bone marrow oedema and cysts varied from 0.73 to 0.82 for experienced raters and from 0.63 to 0.75 for trainees with lowest 95% CI of 0.48 and 0.63. When bone marrow lesions were combined into a total joint score the level of agreement increased to between 0.88 and 0.97 with lowest 95% confidence interval of 0.86. Combining cartilage zone scores did not increase the reliability coefficients.

Conclusion

An expert panel considered that cartilage degradation and subchondral bone marrow lesions where the most important features for staging the severity of ankle OA on MR imaging. Experienced observers can grade the severity of ankle OA on MR with a clinically useful high degree of reproducibility.

Introduction

In the UK, the incidence of symptomatic osteoarthritis (OA) of the ankle has been estimated at 47.7 per 100,000 and is most commonly secondary to trauma. The incidence is increasing and ankle OA is likely to become an increasing health burden[1,2]. The other causes, accounting for 22% of ankles with OA, include rheumatoid arthritis, haemochromatosis, haemophilia, talar dome avascular necrosis and septic arthritis[3,4].

Non-operative treatment for ankle OA includes modified footwear, bracing, oral non-steroidal anti inflammatories, and intra-articular injections of hyaluronic acid or corticosteroids[5]. Failing these patients have a number of operative treatment options including arthroscopy, osteotomies, distraction arthroplasty, ankle arthrodesis and total ankle replacement (TAR)[6]. Tibiotalar arthrodesis is a long established option for end-stage ankle OA that often provides excellent pain relief but at a risk of accelerated arthrosis at the subtalar joint and joints of the midfoot leading to accelerated OA of the surrounding joints[7,8]. To address this patients are increasingly being offered total ankle arthroplasty with third generation implants[9]. While there is evidence of long term positive impact on patients' lives following TAR there is still an annual failure rate and it is difficult to identify which patients are most likely to benefit in either the short or the long term[5]. A number of factors will be considered before offering a patient conservative or operative treatments, which will include quality of life, body mass index, comorbid diseases and radiographic severity of osteoarthritis.

The grading of ankle OA on conventional radiographs is usually performed by measuring minimum joint space width on standing views or by using the Kellgren-

Lawrence (KL) system that produces a score of 0 to 4[10]. MR imaging offers the potential for developing a more sophisticated score of the severity of OA by including features that cannot be demonstrated directly on conventional radiographs, such subchondral bone marrow lesions (BML) and direct evaluation of articular cartilage, as has already been done in the knee[11–15] and hip[16,17]. This might allow for more accurate phenotyping of ankle OA that could allow for better selection of patients into therapeutic pathways.

The aim of this study was to define an MR scoring system for the assessment of ankle OA and to test its inter and intra-rater reliability.

Materials and methods

Local Research Ethics Committee approval was obtained for this retrospective reliability study.

Patient selection

MR examinations of the ankle were chosen sequentially from our institution's Picture Archiving and Communication System (PACS) if there was a preceding ankle radiograph within 4 months of the MR examination and patients were over the age of 18. Exclusion criteria included any history of inflammatory arthritis, previous surgery to the ankle, recent trauma, bone tumour in that limb, haemoglobinopathy, haemachromatosis or any neurological condition limiting function e.g. hemiplegia following stroke.

The ankle radiographs were consensus scored by two radiology trainees with two and four years' experience in reporting appendicular radiographs (SHA and SL) using the modified Kellgren Lawrence score[10]. Cases that met the inclusion criteria were included until there were 10 examinations in each of the five Kellgren-Lawrence groups (n=50). Cases were then assigned a unique identifier code, anonymised and sent back to PACS in an anonymised format with only the unique identifier code present on the MR examination. Radiologists did not have access to the radiograph while grading the MRI.

MR Imaging ankle protocol

MR examinations were performed on either a 1.5T or 3T MR machine (GE Healthcare Systems). The standard protocol included T1 weighted TSE sagittal, T2 weighted TSE fat suppressed sagittal, proton density weighted coronal, T2 weighted STIR coronal and T2 weighted TSE axial sequences. Patients were only eligible for inclusion if the full protocol was completed.

Sample Size

A sample size of 50 was selected based on sample size calculations, considerations regarding underlying marginal prevalence of disease and feasibility. Tables outlined by Sim et al and nomograms outlined by Hong et al were used with the assumption of an underlying equal marginal prevalence of disease[18,19]. The sample of 50 allowed an equal number of 10 examinations for each Kellgren-Lawrence grade zero to four to be used for MR scoring.

For osteophytes, to detect a kappa of 0.61 with H_0 set at 0.2 a sample of n=31 is required. For the zonal assessment of bone marrow lesions (BML), bone marrow oedema (BMO), cysts and cartilage a sample of n=53 is required to detect a kappa of 0.61 with H_0 set at 0.4. For total joint scores of BML, BMO, cysts and cartilage a sample size of n=41 is required to detect a kappa of 0.81 with H_0 set at 0.6. A sample size of 50 examinations was therefore considered appropriate.

Raters

Reader 1 and 2 were radiology trainees (SL and SHA) with four and two years' experience respectively. Readers 3 and 4 were two consultant musculoskeletal radiologists each with more than ten years' experience in reporting musculoskeletal examinations (AT and JC).

MRI Scoring

The MR features used in this study were identified through a Delphi survey of an expert panel comprising musculoskeletal radiologists, rheumatologists, and foot and ankle surgeons (Supplementary file, Table 1). All raters performed the MRI scoring. Each reader was provided with written descriptors for each grade for each variable. No test sets or atlases were used. Inter-rater reliability was assessed between the two experienced radiology consultants and between the two radiology trainees. Intra-rater reliability was tested on a sample of ten MRI examinations that reflected an equal spread across KL grades and were randomly reordered for the second read using an online random number generator (www.random.org). The second read was performed by all raters at least 4 weeks after the first read.

MR grading system

The ankle joint was divided into 16 zones with each variable, except osteophytes, scored in each subregion. The talar dome was divided into nine equal zones by way of a three-column by three-row grid as outlined by Raiken[20]. The nine equal zones were assigned numerical identifiers from one to nine beginning with the most anterior and medial region, proceeding laterally, then posteriorly (Figure 1). The medial and lateral

aspects of the talus were labelled zones 10 and 11 respectively. There was no further subdivision from anterior to posterior of these zones. The distal tibial articulation is divided into three zones from medial to lateral representing zones 12 to 14. Zone 12 therefore articulates with zones 1, 4 and 7 of the talus. Zone 13 articulates with zones 2, 5 and 8. Zone 14 articulates with zones 3, 6 and 9. The medial malleolus represents zone 15, adjacent to zone 10 of the talus. The medial aspect of the distal fibula represents zone 16, adjacent to zone 11 of the talus. The raters identified each zone by eye using the diagram in Figure 1 as a guide.

Bone marrow lesions, bone marrow oedema and subarticular cysts were scored in each of the 16 zones. Bone marrow lesions were defined as any subchondral fluid-signal abnormality and therefore included both subchondral cysts and bone marrow oedema. Bone marrow oedema was described as an ill-defined subchondral hyperintense signal on fluid sensitive sequences. Subchondral cysts were described as well-defined areas of high signal on fluid sensitive sequences (Figures 2 & 3). Osteophytes were recorded as a binary outcome of present or absent.

Cartilage integrity was graded on a six-point scale with a score recorded for each zone. If a cartilage lesion spanned multiple zones a score was recorded for each zone. The system used for grading cartilage integrity was a modified version of the Noyes system of cartilage grading for MRI (Figures 2 & 4, Supplementary figures 1-3)[21,22].

Statistical Analysis

Inter and intra-rater agreement was calculated using the weighted and unweighted kappa coefficients. All statistical analyses were performed using the R statistical

147 programming language using the base package with the additional “irr”, “psych”, and
148 “ggplot2” packages[23].

149 Results

150 Patient Demographics

151 The sample of 50 patients included 27 men and 23 women with a mean age of 54 years
152 (range 23-83). The mean age increased from a mean of 35 years for KL grade 0 to 69
153 years for KL grade 4.

154 Frequency distribution of disease

155 Combining the first and second set of observations in 50 ankles a total of 3200 scores
156 were performed by the four raters for each of the following MR features: cartilage
157 degeneration, bone marrow oedema and subchondral cysts, in the 16 zones. Of the
158 3200 observations for cartilage degeneration approximately one third (n= 1104) were
159 abnormal (Grade 1-5) and two thirds (n=2196) were normal (Grade 0). Grades 1 to 5
160 made up 34%, 19%, 12%, 30% and 7.5% of the abnormal cartilage scores respectively
161 with an even distribution across all zones (Figure 5). Subchondral bone marrow
162 oedema was recorded in 19% (609) of all zones (median 39, IQR: 28-43) and
163 subchondral cysts were recorded in 4.8% (155) of all zones (median 8, IQR: 5-13). Bone
164 marrow oedema and subchondral cysts were recorded in all zones (minimum n=8 and
165 2 respectively).

MR Grade

Osteophytes

The kappa coefficient of agreement for the presence of osteophytes was 0.64 (95% CI: 0.43, 0.85) for the trainee radiologist raters 1 and 2, and 0.92 (95% CI: 0.81, 1.0) for the experienced radiologist raters 3 and 4 (Table 2). The difference between trainee and experienced raters was similar when intra-rater agreement was measured with kappa coefficients of 0.78 (95% CI: 0.39,1.0) for rater 1 and 2 and perfect agreement of k = 1.0 for raters 3 and 4 (Table 3).

Bone marrow signal abnormality

The inter-rater agreement for subchondral bone signal abnormalities varied from 0.63 to 0.75 for raters 1 and 2, and 0.73 to 0.82 for raters 3 and 4, with the lowest agreement for subchondral cysts. The lowest 95% confidence limits were 0.48 and 0.63 respectively. These kappa values were the result of considering the score for each zone separately. When the scores were combined to produce a sum for each feature: bone marrow lesion, bone marrow oedema and subchondral cyst this lead to an increased level of agreement with kappa coefficients for all raters of between 0.88 and 0.97 and a lowest limit of agreement of 0.86 (Table 2).

A similar effect was demonstrated with measures of intra-rater reliability which varied from 0.62 to 0.89 for the trainee radiologist raters 1 and 2, and from 0.51 to 0.85 for the experienced radiologist raters 3 and 4 with lower 95% CI of 0.45 and 0.19 respectively when scores from all zones were considered separately. When the scores

were combined all kappa coefficients increased to a range from 0.77 to 0.97 with ten of the twelve of the lower 95% confidence intervals being 0.76 or greater (Table 3).

Cartilage

The inter-rater kappa coefficient of agreement for raters 1 and 2 was 0.71 (95% CI: 0.67, 0.76) and for raters 3 and 4 was 0.88 (95% CI: 0.85, 0.91) with increases to k = 0.88 and k = 0.96 respectively when the cartilage score for all zones was summated (Table 2). Intra-rater agreement ranged from 0.82 to 0.88 with a lowest 95% CI of 0.7 for all zones considered separately and this increased to for three out of four raters to between 0.81 and 0.95 but with a drop in the lower bound of the 95% CI to 0.4 (Table 3).

When the nine zones of the talar dome were combined into three adjacent strips, running lengthwise from anterior to posterior, that matched the three zones on the tibial plafond there was a slight increase in the kappa coefficient of agreement for the raters 1 and 2 from 0.74 to 0.79, and an equivalent decrease for the raters 3 and 4 from 0.89 to 0.85 (Table 4).

For individual zones 10 to 16 (excluding the talar dome) the kappa coefficient for inter-rater agreement varied from 0.31 to 0.85 for raters 1 and 2, and from 0.76 to 0.92 for raters 3 and 4. There was no one zone that performed consistently worse than any other (Table 4).

Modified cartilage score

Post-hoc kappa coefficients for five alternative cartilage scoring systems were tested for inter-rater agreement. Each system was a simplified version of the modified Noyes

where one or more of the 6 ordinal grades were combined. Two of the systems were 5-point scales, two were 4-point and one system was a 3-point scale. These demonstrated no improvement in the level of agreement with kappa coefficients varying by less than 0.4 from the original scoring system (Table 5).

Surface extent score

Quantification of the extent of cartilage damage was calculated as the number of zones involved per joint. For the first measure all zones with any cartilage damage were included, for the second measure only zones with full thickness cartilage damage were recorded. Inter-rater agreement for all raters assessing any or full thickness cartilage damage varied between 0.93 and 0.95 with a lowest 95% confidence interval of 0.89. Intra-rater agreement for the surface extent of any cartilage damage was similarly high varying from 0.86 to 0.97 but a little lower for full thickness cartilage damage: 0.73 to 1.0.

Discussion

This study provides criteria and reliability data for the staging of osteoarthritis of the ankle. The results suggests that NOAMS does provide a reproducible method for grading the severity of ankle OA on MR imaging. These data apply to grading of OA at a single point in time and do not measure sensitivity to change or test-retest reliability which would be needed for implementing the technique in longitudinal studies.

With possible scores of 0 to 5 for each of 16 zones this grading system can be used to provide a total score from 0 to 80 that describes the total burden of disease in the ankle. However this approach presents a number of questions that have yet to be answered. For instance a total joint score of 10 may indicate either grade 5 osteochondral disease in two zones or grade 2 cartilage disease in 5 zones. While the total joint score is the same this may not correlate with pain or function scores. It may be that it is the most severe osteochondral disease that determines clinical outcomes and not the surface extent.

The granularity of a total joint score with a range of 0 - 80 is at this stage only likely to be useful in research settings where it may be important to detect small changes in MR as an imaging biomarker of OA after a therapeutic intervention. As we come to understand the relative clinical significance of the osteochondral severity, surface extent and anatomical location of disease this MR score can be modified to reflect the importance of these variables. Any modification can then be applied secure in the

knowledge that each element can be scored independently by experienced raters and that simplification of the scoring system does not affect its reliability.

The score can be implemented immediately into clinical practice in a narrative form whereby focal degeneration can be graded from 0 to 5, knowing from previous work that this correlates with arthroscopic findings in the knee, and now that it can reliably assessed in the ankle. Validation against arthroscopic findings would be the next step in assessing the diagnostic accuracy of MR in ankle OA.

The results of this reliability study suggest that the inter-rater reliability of NOAMS is "substantial" to "almost perfect" using the criteria defined by Landis & Koch[24] for the interpretation of kappa statistics, for all measures except for scores relating to individual zones. Intra-rater agreement is similar except for a single score dropping into the "moderate" agreement category. By these criteria the reliability of the NOAMS system for quantifying osteoarthritis of the ankle appears to be at least as reliable as the previously published WORMS[11] and MOAKS[14] systems in the knee and the SHOMRI system in the hip[17]. Although the Landis and Koch categories are widely used they have been criticised because relatively low kappa coefficients of greater than 0.41 are interpreted as "moderate" agreement. This denotes that there is some agreement between raters but it is unlikely to be clinically useful. McHugh suggests a stricter interpretation of kappa where values below 0.6 are classified as "weak" and clinically useful coefficients of agreement $\geqslant 0.8$ as "strong"[25,26]. There is also a view that the reproducibility of clinical studies should be measured from the lower of the 95%

265 confidence intervals because this is the minimum level of agreement that can be
266 confidently assumed from the given sample size[27].

267 The inter and intra-rater reliability for total joint scores for experienced raters was
268 "strong" with just intra-rater kappa for subchondral cysts scoring a "moderate" 0.77
269 (Tables 2 and 3). The lower 95% confidence limit for kappa was "strong" in three, and
270 "moderate" in three out eight comparisons. The less experienced trainee raters did not
271 perform as well on most measures. These results suggest that the inter-rater and intra-
272 rater agreement for NOAMS measures of osteoarthritis can be suitable for clinical use
273 in the hands of experienced observers even when using the strictest criteria for
274 interpreting coefficients of agreement. Observers with more limited experience may
275 not be reliable enough to produce useful measures however the less-experienced
276 raters in our study did not receive any specific training before the study and therefore
277 outcomes might be improved with specific training sets or atlases.

278 In the scoring systems previously discussed the sample sizes varied significantly from
279 n=19 in WORMS and n=20 in MOAKS, to n=109 in the system outlined by Park[11,14,15].
280 No statistical justification for these sample sizes was reported. Sample size calculation
281 for reliability studies is not straightforward with the final sample size a compromise
282 between the power to demonstrate reliability for each variable and what is feasible.
283 The sample size calculation predicted that n=50 would be more than enough to
284 demonstrated a kappa of 0.61 or more for the presence of osteophytes (n=31) and
285 0.81 for total joints scores of bone marrow lesions and cartilage damage (n=41) and
286 this proved to be correct with very narrow 95% confidence limits around these kappa
287 coefficients.

It has been common practice for previously described scoring systems to divide the joint of interest into zones each of which is scored individually depending on the feature being graded. These subdivisions can appear complex and the rationale is not always clear. In the SHOMRI[17] system divides the hip into 10 zones and HOAMS[16] uses nine zones for cartilage and 15 zones for subchondral bone marrow. The knee has been variously divided into the 15 and 14 zones WORMS[11] and MOAKS[14] respectively, and into 9 zones by both the BLOKS[12] and KOSS[13] scoring systems with Park et al. simplifying things further by dividing the knee into 3 regions. The methods are varied and the advantages of one approach over another are not always clear.

The simplest method for describing the position of osteochondral lesions of the talar dome requires just three zones: medial, lateral and central[28] whereas more complex descriptions divide these three zones again forming a 3 x 3 grid of nine zones. The rationale was that most osteochondral lesions occur in the central portion of the medial dome and therefore using a 2 x 2 matrix would leave the most frequently occurring lesions straddling two zones and therefore the 3 x 3 grid was the smallest matrix that would include these lesions in a single zone[20]. Reducing the 3 x 3 grid into 3 zones produced no improvement in inter or intra-rater reliability and therefore there is nothing to be gained by simplifying the 3 x 3 grid. This is useful because the scoring system with the most zones is the most sensitive to change over time.

The methods for grading the severity of each MR feature at each location also varies between systems in other joints. A modified Noyes system has been recommended for use in the ankle[21,29] but without any evidence of reliability or validation. Park et al used a modified Noyes classification for cartilage grading and the grading system used in

KOSS is very similar with only the exclusion of grade 1[13,15]. Only the KOSS scoring system incorporates a specific component for any grading of osteochondral lesions. The Park scoring system classes a grade 3 Noyes as a full thickness cartilage defect with bony involvement therefore including a bony component or osteochondral injury. For simplicity, the presence of bony involvement was classified at the most severe end of the scale of cartilage involvement and this was classified as a grade 4 in this current modified Noyes system.

The modified version of the Noyes score therefore used in this reliability study is a combination of that initially outlined by Recht with the addition of a grade 1 score as proposed by Kijowski[21,22] and a grade 4 to recognise subchondral bone involvement. The results of this study suggest that the reliability of the most detailed modified Noyes system for grading cartilage disease is “strong” and that simplifying the grades does not improve consistency.

The kappa value is influenced by the marginal prevalence of the attribute (the trait prevalence in the study population). The kappa statistic alone is appropriate if the marginal totals are relatively balanced. If the prevalence of given responses is very high or very low the resultant value for kappa may be low even when the observer proportion of reliability is high.

Interpretation of the kappa can be misleading if the marginal prevalences of a particular feature are not relatively balanced causing what is sometimes referred to as the kappa paradox[18,30]. In these circumstances it can be useful to report the percentage agreement alongside kappa which was an option that was adopted in the

MOAKS study when the paradox was suspected. There is no evidence that the kappa paradox has had an influence on the current study and therefore the authors feel that the kappa statistics answer the primary research question without the need for presenting percentage agreement.

There are potential limitations in using trainees to select the patients for this study and in the use of MR machines of different field strengths. While more experienced radiologists might have been more accurate at assigning KL grades to the plain radiographs the aim of this process was to create an even spread of the severity of disease in the study group in order that this did not have an adverse effect on the reliability statistics. For this absolute accuracy is not required as long as the severity of MR scores is evenly distributed and this turned out to be the case (Figure 5). It is possible that by limiting examinations to a single MR machine and a single field strength we could have achieved better reliability but it is reassuring that good reliability can be achieved in a more real-world setting of mixed field strengths.

In conclusion this study describes an grading system for osteoarthritis of the ankle using MR imaging features identified through a multidisciplinary expert panel Delphi survey. This study suggests that experienced observers can grade the severity of ankle OA on MR with a clinically useful high degree of reproducibility.

Conflict of interest

The authors declare no conflict of interest

358 **References**

1. Goldberg AJ, MacGregor A, Dawson J, *et al.* The demand incidence of symptomatic ankle osteoarthritis presenting to foot & ankle surgeons in the United Kingdom. Foot (Edinb) 2012;**22**(3):163–6. https://doi.org/10.1016/j.foot.2012.02.005.
2. Saltzman CL, Salamon ML, Blanchard GM, *et al.* Epidemiology of ankle arthritis: report of a consecutive series of 639 patients from a tertiary orthopaedic center. Iowa Orthop J 2005;**25**:44–6.
3. Valderrabano V, Horisberger M, Russell I, Dougall H, Hintermann B. Etiology of ankle osteoarthritis. Clin Orthop Relat Res 2009;**467**(7):1800–6. https://doi.org/10.1007/s11999-008-0543-6.
4. DiStefano JG, Pinney S. Ankle Arthritis: Etiology and Epidemiology. Seminars in Arthroplasty 2010;**21**(4):218–22. https://doi.org/10.1053/j.sart.2010.09.002.
5. Grunfeld R, Aydogan U, Juliano P. Ankle arthritis: review of diagnosis and operative management. Med Clin North Am 2014;**98**(2):267–89. https://doi.org/10.1016/j.mcna.2013.10.005.
6. Tellisi N, Fragomen AT, Kleinman D, O'Malley MJ, Rozbruch SR. Joint preservation of the osteoarthritic ankle using distraction arthroplasty. Foot Ankle Int 2009;**30**(4):318–25. https://doi.org/10.3113/FAI.2009.0318.
7. Fuchs S, Sandmann C, Skwara A, Chylarecki C. Quality of life 20 years after arthrodesis of the ankle. A study of adjacent joints. J Bone Joint Surg Br 2003;**85**(7):994–8. https://doi.org/10.1302/0301-620x.85b7.13984.
8. Hintermann B. Total Ankle Arthroplasty: Historical Overview, Current Concepts and Future Perspectives. Springer Science & Business Media; 2005.

9. Zaidi R, Cro S, Gurusamy K, *et al.* The outcome of total ankle replacement: a systematic review and meta-analysis. Bone Joint J 2013;**95-B**(11):1500–7. https://doi.org/10.1302/0301-620X.95B11.31633.
10. Kraus VB, Kilfoil TM, Hash TW, *et al.* Atlas of radiographic features of osteoarthritis of the ankle and hindfoot. Osteoarthritis and Cartilage 2015;**23**(12):2059–85. https://doi.org/10.1016/j.joca.2015.08.008.
11. Peterfy CG, Guermazi A, Zaim S, *et al.* Whole-Organ Magnetic Resonance Imaging Score (WORMS) of the knee in osteoarthritis. Osteoarthritis Cartilage 2004;**12**(3):177–90. https://doi.org/10.1016/j.joca.2003.11.003.
12. Hunter DJ, Lo GH, Gale D, Grainger AJ, Guermazi A, Conaghan PG. The reliability of a new scoring system for knee osteoarthritis MRI and the validity of bone marrow lesion assessment: BLOKS (Boston Leeds Osteoarthritis Knee Score). Ann Rheum Dis 2008;**67**(2):206–11. https://doi.org/10.1136/ard.2006.066183.
13. Kornaat PR, Ceulemans RYT, Kroon HM, *et al.* MRI assessment of knee osteoarthritis: Knee Osteoarthritis Scoring System (KOSS)--inter-observer and intra-observer reproducibility of a compartment-based scoring system. Skeletal Radiol 2005;**34**(2):95–102. https://doi.org/10.1007/s00256-004-0828-0.
14. Hunter DJ, Guermazi A, Lo GH, *et al.* Evolution of semiquantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score). Osteoarthritis Cartilage 2011;**19**(8):990–1002. https://doi.org/10.1016/j.joca.2011.05.004.
15. Park H-J, Kim SS, Lee S-Y, *et al.* A practical MRI grading system for osteoarthritis of the knee: association with Kellgren-Lawrence radiographic scores. Eur J Radiol 2013;**82**(1):112–7. https://doi.org/10.1016/j.ejrad.2012.02.023.

16. Roemer FW, Hunter DJ, Winterstein A, *et al.* Hip Osteoarthritis MRI Scoring System (HOAMS): reliability and associations with radiographic and clinical findings. Osteoarthritis Cartilage 2011;**19**(8):946–62. https://doi.org/10.1016/j.joca.2011.04.003.
17. Lee S, Nardo L, Kumar D, *et al.* Scoring hip osteoarthritis with MRI (SHOMRI): A whole joint osteoarthritis evaluation system. J Magn Reson Imaging 2015;**41**(6):1549–57. https://doi.org/10.1002/jmri.24722.
18. Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy 2005;**85**(3):257–68. https://doi.org/10.1093/ptj/85.3.257.
19. Hong H, Choi Y, Hahn S, Park SK, Park B-J. Nomogram for sample size calculation on a straightforward basis for the kappa statistic. Annals of Epidemiology 2014;**24**(9):673–80. https://doi.org/10.1016/j.annepidem.2014.06.097.
20. Raikin SM, Elias I, Zoga AC, Morrison WB, Besser MP, Schweitzer ME. Osteochondral Lesions of the Talus: Localization and Morphologic Data from 424 Patients Using a Novel Anatomical Grid Scheme. Foot Ankle Int 2007;**28**(2):154–61. https://doi.org/10.3113/FAI.2007.0154.
21. Kijowski R, Davis KW, Woods MA, *et al.* Knee joint: comprehensive assessment with 3D isotropic resolution fast spin-echo MR imaging--diagnostic performance compared with that of conventional MR imaging at 3.0 T. Radiology 2009;**252**(2):486–95. https://doi.org/10.1148/radiol.2523090028.
22. Recht MP, Piraino DW, Paletta GA, Schils JP, Belhobek GH. Accuracy of fat-suppressed three-dimensional spoiled gradient-echo FLASH MR imaging in the

detection of patellofemoral articular cartilage abnormalities. Radiology 1996;**198**(1):209–12.

23. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing,. Vienna, Austria.: 2020.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;**33**(1):159–74.
25. Nunnally JC, Bernstein RH. The Assessment of reliability. Psychometric Theory New York: McGraw-Hill, Inc 1994:248–92.
26. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;**22**(3):276–82.
27. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;**15**(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012.
28. McGahan PJ, Pinney SJ. Current concept review: osteochondral lesions of the talus. Foot Ankle Int 2010;**31**(1):90–101. https://doi.org/10.3113/FAI.2010.0090.
29. Saifuddin A. Musculoskeletal MRI. 1 edition. London: CRC Press; 2008.
30. Gwet K. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Statistical Methods for Inter-Rater Reliability Assessment 2002;**1**(6):1–6.

Figure Legends

Figure 1

Diagram illustrating (A) zones 1 to 9 on an axial section through the talar dome and (B) zones 11 to 16 on a coronal section.

Figure 2

Diagrammatic representation of scoring scheme for bone marrow lesions and cartilage disease on MRI of the ankle with individual scores for each feature indicated by the white numeral.

Figure 3

Coronal STIR (A) and sagittal T2W fat saturated MR (B & C) demonstrating focal subchondral bone marrow oedema in zone 13 (A: arrow), a solitary subchondral cyst (arrow) in zone 13 with subchondral bone marrow oedema in zones 12 to 14 (B) and extensive subchondral cyst formation (arrowheads) and bone marrow oedema (C) scored in agreement by the two senior raters.

Figure 4

Coronal PD (A) and sagittal T2W FS MR (B) images demonstrating grade 1 cartilage disease with abnormal hyperintense signal on the T2W fat sat images but an intact articular cartilage surface on the coronal PD (arrows).

Figure 5

Jitter plot demonstrating the distribution of scores of the severity of ankle disease across all 16 zones of the tibiotalar joint. Two-thirds of all zones were scored as normal (0). The remaining one third of scores were distributed across all zones demonstrating that the dataset tested the scoring system for all grades of disease in all parts of the articular surface.

Supplementary Figures

Figure 1

Coronal PD (A) and STIR (B) MR images demonstrating grade 2 cartilage disease with partial thickness loss of hyaline cartilage in segment 4 of the superolateral talar dome (arrows).

Figure 2

Coronal PD (A) and STIR (B) MR images demonstrating grade 3 cartilage disease with full thickness loss of hyaline cartilage in zones 6 and 12 of the medial tibiotalar joint (arrows).

Figure 3

Sagittal T1W (A) and T2W FS (B) demonstrating grade 4 cartilage disease (white arrow) with irregularity of the subchondral plate (black arrow).

401 # Tables

402 **Table 1**

403 Delphi survey results for the Tibiotalar joint.

404

	Round 1			Round 2		
	Mean	Median	SD	Mean	Median	SD
Presence of BMO	4	4	0.71	4.13	4	0.6
Extent of BMO	3.75	3.5	0.83	4.13	4	0.78
Presence of osteophytes	3.75	4	1.09	3.88	4	1.17
Number of osteophytes	2.88	3	1.05	2.75	3	0.97
Cartilage integrity	4.63	5	0.48	4.63	5	0.48
Ligament integrity	3	3	1	2.88	3	1.05
Presence of OCD	4.13	4.5	1.05	4.25	4	0.66
Presence of cysts	4.25	4	0.66	4.38	4	0.48
Presence of bone attrition	4.25	4	0.66	4.38	4	0.48
Severity of bone attrition	4.38	4.5	0.7	4.38	4	0.48
Presence of joint effusion	2.75	3	0.66	3.13	3	1.05
Presence of synovitis	2.88	3	0.6	2.88	2.5	1.05
Presence of loose bodies	2.5	2.5	0.5	-	-	-

BMO; Bone marrow oedema. OCD; Osteochondral defect. SD; Standard deviation

405

406

Table 2

Table demonstrating the kappa coefficients of inter-rater agreement with 95% confidence intervals for bone marrow signal abnormalities considered separately for each zone and with the sum of scores for all zones affected.

	Raters	
Feature	1 and 2	3 and 4
Osteophytes	0.64 (0.43,0.85)	0.92 (0.81,1.00)
All zones		
Bone marrow lesion	0.75 (0.69,0.81)	0.82 (0.77,0.87)
Bone marrow oedema	0.73 (0.67,0.79)	0.81 (0.75,0.86)
Subchondral cysts	0.63 (0.48,0.78)	0.73 (0.63,0.83)
Cartilage	0.71 (0.67,0.76)	0.88 (0.85,0.91)
Sum of all zones		
Bone marrow lesion	0.97 (0.96,0.99)	0.96 (0.93,0.98)
Bone marrow oedema	0.97 (0.97,0.99)	0.94 (0.90,0.98)
Subchondral cysts	0.94 (0.87,1.00)	0.91 (0.86,0.97)
Cartilage	0.88 (0.82,0.95)	0.96 (0.92,0.99)

Inter-rater weighted kappa. $p < 0.01$

Table 3

Table demonstrating the kappa coefficients of intra-rater agreement with 95% confidence intervals for bone marrow signal abnormalities considered separately for each zone and with the summated score for all zones affected.

Feature	Reader 1	Reader 2	Reader 3	Reader 4
Osteophytes	0.78 (0.39,1.0)	0.78 (0.39,1.0)	1.0	1.0
All zones separate				
Bone marrow lesion	0.73 (0.60,0.85)	0.60 (0.45,0.76)	0.85 (0.75,0.96)	0.70 (0.56,0.83)
Bone marrow oedema	0.69 (0.56,0.82)	0.62 (0.47,0.77)	0.83 (0.72,0.94)	0.70 (0.56,0.83)
Subchondral cysts	0.89 (0.75,1.00)	0.79 (0.52,1.00)	0.81 (0.61,1.00)	0.51 (0.19,0.83)
Cartilage damage	0.85 (0.70 0.91)	0.82 (0.75 0.90)	0.88 (0.81 0.92)	0.84 (0.77 0.91)
Sum of all zones				
Bone marrow lesion	0.97 (0.94,0.99)	0.79 (0.67,0.92)	0.94 (0.85,1.00)	0.91 (0.78,1.00)
Bone marrow oedema	0.97 (0.94,0.99)	0.93 (0.83,1.00)	0.92 (0.83,1.00)	0.91 (0.78,1.00)
Subchondral cysts	0.91 (0.77,1.00)	0.88 (0.76,1.00)	0.97 (0.89,1.00)	0.77 (0.50,1.00)
Cartilage damage	0.95 (0.91 0.98)	0.76 (0.40,1.00)	0.81 (0.59,1.00)	0.94 (0.89 0.99)

Intra-rater. Weighted kappa. $p<0.05$

Table 4

Inter-rater kappa coefficients of agreement for scoring of cartilage disease.

Location	Raters 1 and 2	Raters 3 and 4
Talar Dome 9 zone	0.74 (0.68,0.80)	0.89 (0.86,0.92)
Talar Dome 3 zone	0.79 (0.72,0.86)	0.85 (0.79,0.91)
Zone 10	0.66 (0.45,0.86)	0.87 (0.79,0.95)
Zone 11	0.31 (0.07,0.55)	0.91 (0.84,0.98)
Zone 12	0.60 (0.37,0.84)	0.76 (0.61,0.92)
Zone 13	0.77 (0.63,0.91)	0.92 (0.82,1.00)
Zone 14	0.85 (0.76,0.94)	0.91 (0.84,0.97)
Zone 15	0.45 (0.21,0.69)	0.76 (0.56,0.96)
Zone 16	0.45 (0.16,0.74)	0.91 (0.82,0.99)

Inter-rater weighted kappa. $p<0.01$

Table 5

Inter-rater agreement coefficients for each of the simplified versions of the NOAMS score demonstrating little improvement in agreement with fewer increments in each feature score.

Version	Raters 1 & 2	Raters 3 & 4
Original	0.71 (0.67, 0.76)	0.88 (0.85,0.91)
1	0.70 (0.65, 0.74)	0.85 (0.82, 0.88)
2	0.70 (0.66, 0.75)	0.85 (0.82, 0.88)
3	0.65 (0.59, 0.71)	0.85 (0.82, 0.88)
4	0.67 (0.61, 0.73)	0.85 (0.82, 0.88)
5	0.72 (0.67, 0.76)	0.88 (0.85, 0.91)

Inter-rater weighted kappa. $p<0.01$