Semi-Quantitative Analysis of Magnetic Resonance Imaging in Arthritis: The Pursuit of Optimal Scoring Granularity

by

Stephanie Anne Wichuk

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Medical Sciences - Radiology and Diagnostic Imaging

University of Alberta

Abstract

Semi-quantitative image scoring systems have been developed to assess various forms of arthritis through assignment of numerical scores corresponding to the extent of lesion presence within an image. Such scoring methodologies allow researchers to systematically record differences between patients and within the same patient over time, allowing for detailed analyses of imaging findings as they relate to disease severity, treatment response, and various clinical markers. The level of detail, or granularity, contained within a scoring system influences the ability of the score to convey smaller differences or changes in lesion extent between images, but can also impact the reliability of scores assigned to the same images by different readers as well as the amount of time and effort required of readers. This thesis examines how scoring granularity impacts the utility of semi-quantitative image data, as well as inter-rater reliability.

Chapter 1 contains a study of inflammatory and structural lesion scoring in the sacroiliac joint of ankylosing spondylitis patients at different levels of granularity, showing how analysis of semi-quantitative scoring data can produce different results depending on which reader's data is used.

Chapters 2 and 3 both deal in semi-quantitative scoring of bone marrow lesions in knee osteoarthritis patients. Chapter 2 demonstrates how reliability

of scoring can–but does not always– decrease as the regions of interest decrease in size, while chapter 3 covers an extensive analysis of bone marrow lesions at different levels of detailed scoring, and explores the use of an artificial intelligence-generated semi-quantitative scoring output to test these results in a larger dataset. This final chapter suggests a justification for scoring in granular detail, and makes a case for the need to ensure training of artificial intelligence algorithms is targeted for reliable detailed scoring.

Preface

Chapter 1 of this thesis is in the revision stage with Acta Radiologica, and, assuming acceptance, will be published as “Wichuk S, et al. “Algorithm-generated sacroiliac joint MRI predictors of axial Spondyloarthritis: An analysis of the Assessment of SpondyloArthritis International Society MRImagine Study dataset” in Acta Radiologica 2024. S. Wichuk conceptualized the study, performed the data analysis, and wrote the manuscript with the input of the listed co-authors.

Chapters 2 and 3 are intended to be individual manuscripts for publication in an appropriate rheumatology journal, but have not yet reached the submission stage at the time of thesis submission. Again, S. Wichuk conceptualized these studies, performed the data analysis, and completed all writing.

All other contents of the thesis is original unpublished work by S. Wichuk.

Dedication

For Arthur and Leo. Thank you for bearing with me.

Acknowledgements

I would like to thank Dr. Jacob Jaremko for taking me on as his graduate student and patiently guiding me through the process of turning my vague ideas into something more concrete. I would also like to thank Drs. Abhilash Hareendranathan and Yan Yuan for their continual support and insight as members of my supervisory committee. Additionally, this work would not be possible without the up front time and effort of each expert reader who provided scoring data to the ASAS MRImagine and OMERACT Knee reading projects and Dr. Banafshe Felfeliyan who developed the iKIMRISS BML algorithm used in the final chapter.

Finally, I must express my sincere gratitude to Dr. Walter Maksymowych who mentored me for well over a decade prior to my becoming a graduate student, affording me myriad opportunities for growth and steering me toward this present juncture. Without his influence, I would never have been introduced to the topic at hand.

Table of Contents

List of Tables

List of Figures

List of Abbreviations:

- CHAID: Chi-Square Automated Detection Algorithm
- SIJ: Sacroiliac Joint
- BME: Bone Marrow Edema
- BML: Bone Marrow Lesion
- OR: Odds Ratio
- CI: Confidence Interval
- SD: Standard Deviation
- ICC: Intra-Class Correlation Coefficient
- WOMAC: Western Ontario and McMaster Universities Arthritis Index
- KIMRISS:Knee inflammation MRI Scoring System
- OAI:Osteoarthritis Initiative
- MOAKS: MRI Osteoarthritis Knee Score

INTRODUCTION

What can we learn from images?

Both forms of arthritis discussed in this work-–osteoarthritis of the knee (OA) and axial spondyloarthritis(axSpA)–-have varying presentations in the clinic and varying implications for a patient's quality of life and prognosis. Factors such as phenotype [1,2] and current stage of disease [3,4] can influence the relative ease of a physician's diagnostic and prognostic determinations [5,6] as well as selection of appropriate treatment plans. While patient history, clinical assessments of mobility [7,8], measurements of certain laboratory markers[9], and self-reporting of symptoms and health-related functioning in daily life[10,11] are integral to the picture of a patient's diagnosis and current disease state, these can lack either sensitivity or specificity for diagnostic and prognostic variables, even when considered together.

Various medical imaging protocols in arthritis can aid in developing a clearer snapshot of current disease activity and progression over time, bringing to light aspects of pathology that have not yet manifested clinically or have undergone change between clinic visits. Plain radiographs (x-rays) are a crucial factor in diagnosis and serve as the hallmark of structural disease progression in both OA and axSpA[12–14]. Cartilage loss and joint space narrowing are detected and monitored using using knee radiographs, for example, while

x-rays of the pelvis and spine can display the sclerotic changes (hardening of subchondral bone), joint space narrowing, and eventual fusion of the sacroiliac joints and vertebrae associated with axial spondyloarthritis.

Although x-rays are relatively quick and cost-effective to obtain, and are used extensively in patient care and as research endpoints, they are 2-dimensional and nonideal for visualizing certain aspects of the anatomy–for example, soft tissue and structures obstructed by other structures in the plane of view. They are also insensitive to disease progression over shorter periods of time [15,16]. More recently developed magnetic resonance imaging (MRI) techniques help overcome some of these issues, shedding light on the pathophysiology of both diseases and allowing for earlier lesion detection and observation of smaller longitudinal changes compared to plain radiographs [17,18]. Structures are visualized more completely through many cross-sectional slices, providing a robust representation of the size and location of lesions. The variety of lesions that can be observed on MRI also far exceeds what can be discerned from x-ray. Fluid-sensitive sequences can be evaluated for inflammatory lesions, while fine structural details are sharply visualized on T1W sequences, allowing for detection of abnormalities such as small erosions of the cortical bone and infiltration of fatty tissue into the subchondral bone.

Semi-quantitative scoring

Given the treasure trove of information contained within these images, it is pragmatic to collect and analyze imaging data to investigate its relationships

to diagnosis, prognosis, and severity of symptoms and impairment. Large scale analyses of qualitative findings, such as those derived from a radiologist's report, are challenging due to the need for conversion into standardized data. This conversion requires nuanced but consistent interpretation of text by a data custodian, which can be especially difficult due to a lack of standardized language amongst radiologists [19,20]. Furthermore, standard-of-care reporting is not done with a research end-use in mind, and may not explicitly cover the data points of interest for research and clinical trials.

Semi-quantitative scoring methodologies aim to navigate such problems by allowing researchers to record and interpret findings via carefully developed operational definitions of lesion appearance, location, and size[1]. The resulting numerical scores are indicative of the extent of disease pathology across one or several domains.

In addition to eliminating any subjective interpretation of the imaging report by an end-user, these scoring systems have the potential to minimize the subjectivity of an image reader's own interpretations by explicitly defining what they should be looking for and how their observations should be recorded. Figure 1 shows an example of an online semi-quantitative scoring interface for spinal lesions typical of axSpA, available to readers upon the completion of an extensive training module covering lesion definitions, scoring locations, and

[1] Although this work focuses on semi-quantitative scoring systems developed for research and clinical trials in arthritis, image scoring systems have also been developed to standardize image interpretation in a clinical setting. These include the Reporting and Data Systems (RADS) for breast, thyroid, liver, and more.

descriptions of potential error sources. Additionally, an example of interactive training material for sacroiliac joint scoring calibration can be seen in Figure 2.

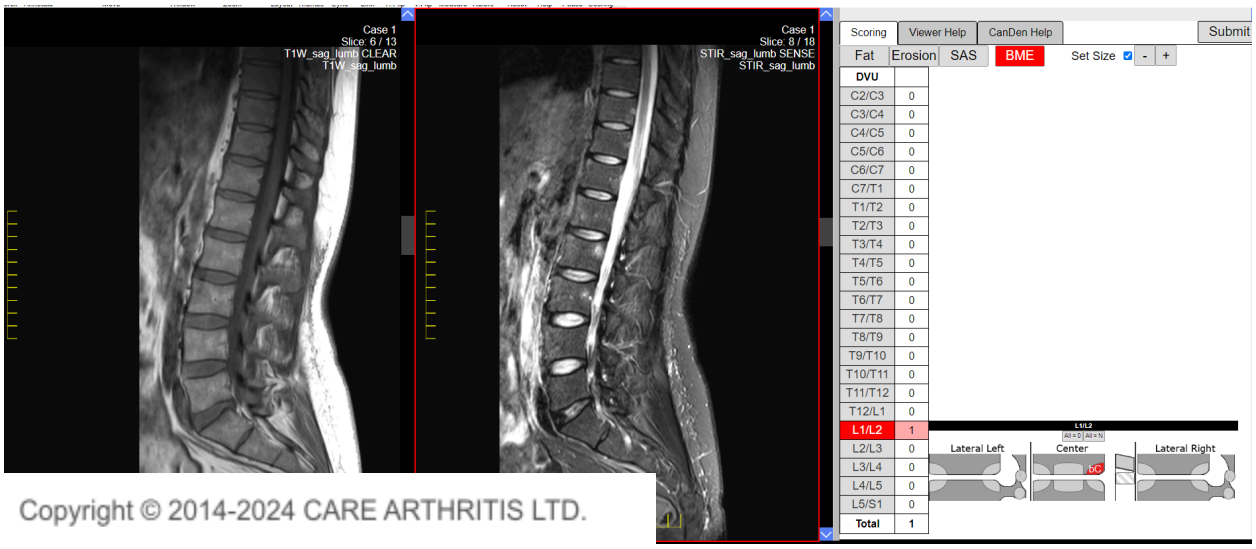


Figure 1. Example of semi-quantitative scoring of spinal lesions (CanDen system) in axial spondyloarthritis MRI

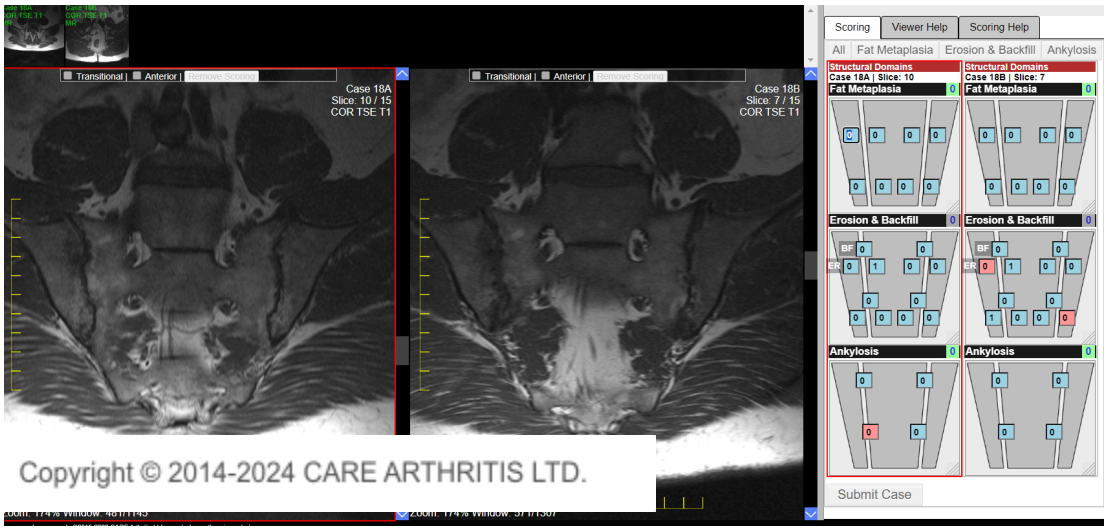


Figure 2. Screenshot from training module for structural lesion scoring on sacroiliac joint MRI according to the Spondyloarthritis Research Consortium of Canada (SPARCC) Structural SIJ Score (SSS). The calibration module provides

real-time feedback to the user, indicating whether their assigned scores are considered correct according to the methodology.

Scoring granularity

There are many approaches to scoring methodology development, and when interpreting the resultant data, it is important to consider how the score has been derived and exactly what it is measuring[21]. Depending on the methodology, scores may be assigned at different levels of granularity (detail). This can apply when distinguishing between lesion subtypes (eg. a system where all bone marrow edema receives the same score vs. one where "intense" bone marrow edema receives an additional score) or size of scoring regions (eg. a system that assigns one dichotomous score for lesion presence within the entire knee joint vs. a system that incorporates dichotomous scoring for lesion presence within many subdivisions of the joint). As the approach to scoring becomes more granular, there is an increased potential to extract finer differences both between cases and over time within the same case.

In the knee bone marrow lesion (BML) scoring templates shown in the left and center frames of Figure 3, the methodologically-defined regions are quite broad and rely on the reader's evaluation of the percentage of each region occupied by a BML. Their estimate of this percentage is recorded in wide-ranging categories rather than on a continuous scale (<10%, 10-25%, >25%; and <33%, 33-66%, and >66% for the scoring systems shown in the left

and center, respectively) [22,23], and as a result, lesions of considerable variation in size may receive the same score.

The frame on the right shows smaller, more numerous scoring regions defined by the Knee MRI Inflammation Scoring System (KIMRISS) that are each meant to be scored dichotomously for BML presence in that region.[24] Rather than relying on the reader's estimate of size, a BML score is based on the number of these small regions it occupies. This should theoretically allow for more consistent detection of small differences, while also providing additional easily interpretable information about where lesions are located. For example, while a score indicating a bone marrow lesion in "<10% of the weight-bearing region" only tells us there is a small focal lesion somewhere in that region, a score in the KIMRISS region labeled "FC1" tells us it is in the anterior part of the region, immediately adjacent to the tibiofemoral joint.

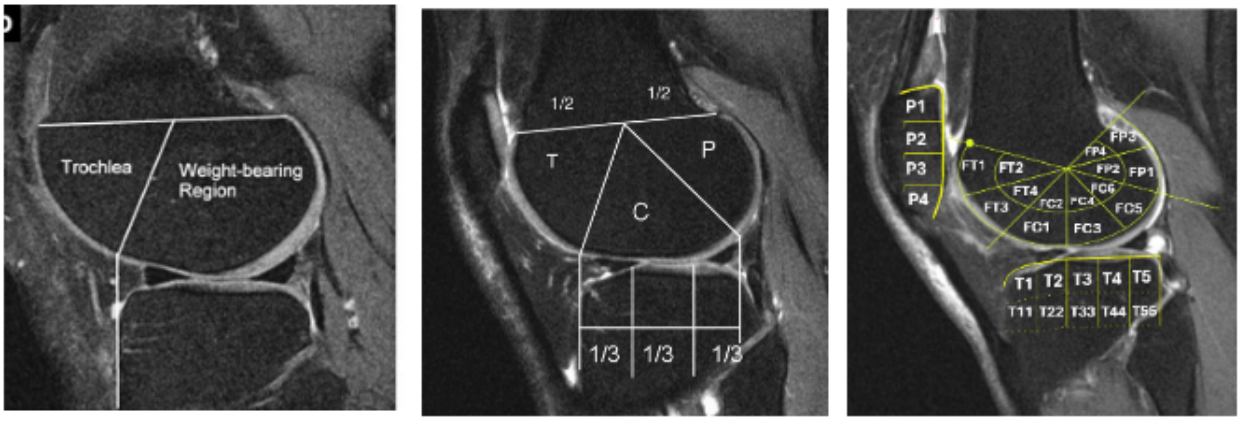


Figure 3. Bone marrow lesion (BML) scoring regions in the femur and tibia as defined by Whole Organ MR Score (WORMS) (left), MRI Osteoarthritis Knee Score (MOAKS)(center), and the Knee Inflammation MRI Score (KIMRISS) (right).

Utility and Feasibility

Scoring interfaces have progressed over time from tedious Excel spreadsheet data entry to online modules complete with ready-made scoring region template overlays and one-click score assignment. However, the time required of an expert reader (often a radiologist) to learn the rules of a system, calibrate, and input scoring remains nontrivial and tends to increase alongside the level of scoring detail required. When optimizing a semi-quantitative scoring system for the level of granular detail captured, the utility of increased detail over a lesser degree of detail is an important consideration. If analyses show that very detailed scoring of lesion extent in small increments across small anatomical subdivisions does not provide any additional insight into relationships to disease state or outcomes when compared to a “yes or no” assessment of lesion presence across an entire scan, the former may not be worth the added time and energy expenditure.

Reliability

In addition to utility, the level of detail captured by a scoring system can also affect reliability between different readers. Although semi-quantitative methodologies aim to minimize subjectivity of scoring, the number of discrete observations that must be recorded in a single case increases with scoring

granularity, producing more chances for discrepancy between different observers. An overall score derived from many observations recorded through a more granular framework can show improved reliability compared to a score that combines fewer discrete observations [25]. While sources of discrepancy may "come out in the wash" when all granular components are combined, they may have more impact when granular components are analyzed separately. The finer distinctions created by more detailed scoring can only be taken at face value if readers demonstrate consistent agreement on these fine distinctions. If there is a demonstrated tendency toward reader disagreement, data generated by the system may contain a certain amount of noise or error and must be analyzed with this in mind.

Technological advances

There has been continual advancement in artificial intelligence (AI) algorithms trained to detect disease pathology from imaging, which, over time, may eliminate the need for human-generated semi-quantitative scoring altogether[26–29]. AI can be tailored to extract data of interest, free from the constraint of expert time and energy required to read images. However, optimal granularity still must be sought during algorithm development, focusing on reliability with human observers to ensure a high signal-to-noise ratio in the data output. Utility also continues to be an important consideration, as fine-tuning algorithms to produce reliable data at a high level of granularity is likely to require significant time and attention. It is important that

development efforts are directed toward obtaining data that provides useful information.

It is worth noting that, with specific regard to arthritis and other rheumatic diseases, the Outcome Measures in Rheumatology (OMERACT) organization has systematically developed criteria for choosing imaging measurement instruments, including semi-quantitative scores, for research and clinical trials[21,30]. To be widely accepted as valid, measurements produced by artificial intelligence will be expected to pass OMERACT's Core Instrument Set Filter, which comprises three arms: 1)Truth (does the measurement really represent what it is supposed to measure), 2)Discrimination (can the measurement discern between groups of interest or detect longitudinal changes?), and 3) Feasibility (is obtaining the measurement practical?). Utility and reliability, referenced above, fall under the Discrimination and Truth arms, respectively.

The following three manuscripts each deal with semi-quantitative arthritis imaging data in a way that centers the question of granularity and its relationship to either utility, reliability, or both. Chapter 1 deals with total scores that result from scoring input made at different levels of granularity, showing that sum scores derived from a more granular scoring framework increases the consistency of diagnostic predictors derived from different readers' data. Chapter 2 is an in-depth reliability analysis of individual components of a granular scoring framework, demonstrating differences in

reliability of individual observations depending on anatomical locations. Chapter 3 explores the relationship of individual granular components of a knee scoring system to clinical outcomes, showing that finer distinctions between cases may be more predictive than differences between broad assessments. It also examines reliability between artificial intelligence and human-generated scores.

CHAPTER 1

Title: Algorithm-generated sacroiliac joint MRI predictors of axial Spondyloarthritis: An analysis of the Assessment of SpondyloArthritis International Society MRImagine Study dataset

Abstract

Background: The Assessment of SpondyloArthritis international Society (ASAS) MRImagine study provides global and semi-quantitative scoring of 12 types of active and structural sacroiliac joint (SIJ) MRI lesions, plus additional diagnostic and clinical data from 135 patients presenting with symptoms of axial spondyloarthritis (axSpA). Some SIJ MRI lesion types or combinations thereof may be more strongly tied to a positive axSpA diagnosis than others. **Purpose:** To use Chi-Squared Automatic Interaction Detection Algorithm (CHAID) to identify combinations of MRI lesions from a multi-feature image scoring system that are strongly predictive of a positive axSpA diagnosis and analyze whether predictor selection is affected by level of scoring granularity. **Methods:** We applied CHAID and traditional logistic regression techniques to active and structural MRI lesion data using axSpA diagnosis by a rheumatologist as the outcome variable. **Results:** Subchondral bone marrow edema (BME) was the most significant predictor in CHAID decision-trees generated at all levels of granularity [OR=13.75 (3.13-60.35) for detailed mean reader scores]. In the absence of BME, presence of fat lesions served as a strong predictor of axSpA at the lowest level of granularity.[OR (95% CI) 6.38

(2.56-30.96) for majority reader agreement] . The strength of odds ratios for diagnosis based on the predictors selected by CHAID generally increased with increased scoring granularity. **Conclusion:** Our findings highlight the strength of the relationship of BME and fat lesions on MRI to rheumatologist's diagnosis of SpA and consistencies in scoring among calibrated experts. The CHAID algorithm can be helpful in automating analysis of relationships between lesion types and diagnosis within complex multivariable image scoring system data sets across different levels of granularity.

Keywords: MR-Imaging, Outcomes Analysis, Joints, Observer Performance, Statistics, Skeletal-Axial

Introduction

The Chi-Squared Automatic Interaction Detection (CHAID) algorithm is a statistical method which automatically distills the most important predictors, and combinations thereof, from datasets with many potentially predictive variables. (1) When optimized using high-quality training and validation datasets, the result may inform clinical decision-making through a visual flowchart of important factors that lead to a particular diagnostic or prognostic odds ratio (OR).

Over the past 15 years, multiple diagnostic imaging studies have applied CHAID to identify individual or combinations of image features most predictive

of a specified diagnostic outcome. For example, a group of researchers have published several studies using results of a CHAID model of 17 breast MRI descriptors to create risk categories for various predefined outcomes([31]), potentially simplifying estimation of patient risk and adding clarity to clinical decision-making processes. Use of CHAID-generated decision trees has also been suggested in MRI liver tumour evaluation([32]), MRI axial skeleton tumour evaluation([33]), osteoporosis screening via dental radiographs([34]), and more.

The approach is best utilized where comprehensive descriptive or semi-quantitative image evaluation data are available in combination with an external diagnostic gold standard or outcome measure of interest, on a sample large enough to allow for robust sample sizes in subcategories following the initial split of the tree.

When many potential predictors exist and may be significant in particular combinations with one another, an algorithm such as CHAID circumvents the need to manually build an unwieldy exploratory multiple regression model factoring in every imaginable interaction between variables. At the same time, in contrast to more advanced “black box” machine learning techniques increasingly utilized in outcome prediction, CHAID provides transparency in predictor selection, creating results that are immediately accessible and translatable to a human observer. CHAID runs chi-squared analyses between each independent variable entered into the model and the

chosen dependent variable, and first splits the sample into two or more categorical groups according to the strongest predictive variable found (i.e. lowest significant p-value after automatic correction for multiple tests). It then repeats chi-squared analyses of remaining independent variables on the subset of cases within each of the categorical groups formed in the initial split to find the variable with most significant predictive power therein. It again splits the cases into further subgroups according to the most significant variable and cut-off value identified, and repeats until no further significant predictors are found within subcategories formed in the previous step. The final result is a decision tree with at least two terminal nodes or "leaves" (Figure 1). Each of these terminal nodes possesses a distinct OR for the outcome given the independent variable value or combination of independent variable values it represents. The software used in this study allows for the inclusion of both categorical and continuous variables, automatically splitting the values of the latter into distinct categorical bins, or groups separated by cut-off points, prior to running the analysis. The number of bins included may be pre-defined by the user based on the nature of the data.

Objective

We performed CHAID analysis on the Assessment of SpondyloArthritis international Society (ASAS) MRImagine study sacroiliac joint MRI dataset, seeking to identify strongly predictive combinations of MRI lesions from this

multi-feature image scoring system, and to analyze whether predictor selection is affected by level of scoring granularity.

Methods

Data Available

ASAS MRImagine study imaging data were derived from an international multi-center cohort of patients with chronic back pain of unknown origin beginning prior to age 45 and symptoms considered suspicious for axial spondyloarthritis (axSpA) [mean (SD) symptom duration 7.4 (7.5) years]. Global and semi-quantitative MRI scoring of the sacroiliac joint (SIJ) was completed by 7 expert readers from the ASAS MRI Working Group for different types of active and structural lesions indicative of axSpA, using semi-coronal STIR and T1 scans. Complete STIR and T1 sequences were available in DICOM format on 135 cases, from which active and structural lesions observed in axSpA were scored dichotomously in an electronic case report form for global presence on the scan (yes/no), overall presence on the sacral and iliac side of both left and right SIJs (yes/no for sacral and iliac portions of each joint) without regard to number of slices affected, and semi-quantitatively for the sum of lesions present throughout the entire readable portion of the scan (slices with at least 1cm of vertical height visible). At the semi-quantitative level of granularity, lesion presence was recorded on each scorable slice in upper and lower portions of each joint (i.e. halves) for fat metaplasia in joint space (backfill) and ankylosis, and upper and lower portions of both sacral and iliac side of each

joint on each slice (i.e. quadrants) for all other lesions. Active and structural lesion scores were assigned according to definitions standardized by international consensus of experts from the ASAS MRI Working Group ([35]). In global scoring, images were evaluated for 5 types of active lesions [subchondral bone marrow edema (BME), capsulitis, inflammation in an erosion cavity, enthesitis outside the SIJ, and joint space fluid], and 7 types of structural lesions (erosion, subchondral fat lesion (fat), subchondral fat lesion >1cm in width (fat >1cm), fat metaplasia in erosion cavity, sclerosis, ankylosis, and bone bud). Detailed semi-quantitative scores were recorded for all aforementioned structural lesion types, but only subchondral BME in the active category. The study also recorded an independent rheumatologist's diagnosis of axSpA (yes/no) for each case, assigned with consideration of clinical workup, laboratory, and imaging findings, without access to any MRI scoring data from the central study readers. Clinical and demographic data were retained for research purposes, but in light of the small sample size relative to the number of MRI variables of interest available, these were not included in the models for this preliminary study to avoid overfitting.

Statistical Methods

Using Medcalc v. 20.027 and IBM SPSS v. 28.0., CHAID analyses were performed on SIJ MRI scoring data of 135 cases in the ASAS MRImagine cohort with detailed active and structural scoring data available, using rheumatologist's diagnosis as the dependent variable. The minimum numbers

of cases required for parent and child nodes of the CHAID tree (i.e. categories allowed to be split into subcategories and resulting subcategories) were set to 10 and 5, respectively, and each tree was validated using the software's built-in 10-fold cross-validation feature. Overarching model versions were created at three different levels of granularity using available independent variables as follows:

Model 1) Global presence of each of the 12 active and structural variables; yes/no for entire scan

Model 2) 2-Dimensional semi-quantitative scoring; number of joint sides (range 0-4) or joints (for joint space enhancement, backfill and ankylosis only, range 0-2) positive for each of the 12 active and structural variables, without regard to number of MRI slices affected.

Model 3) 3-Dimensional semi-quantitative scores for each of the 8 active and structural variables for which this data was recorded, summing up all lesions in each quadrant in all slices affected.

CHAID was performed on all 3 models using individual and majority reader data ($\geq$ 4 readers agree yes/no for model 1, and 7-reader mean score for models 2 and 3) producing a total of 24 decision trees. An example of software output is provided in Figure 1. Odds Ratios were calculated for each node of each decision tree to indicate the odds of positive diagnosis in the presence or absence of the indicated lesion or combination of lesions, in cases of multi-level

trees. The plain software output was also converted to a pictographic representation of the specific MRI lesions demonstrating predictive utility

In order to discern whether CHAID provides any insight over and above more traditional statistical techniques, uni- and multi-variable logistic regression analyses were also performed.

Results

Of 135 cases analyzed, 97 (71.9%) received a positive diagnosis for axSpA. Subchondral BME was the most commonly scored MRI lesion in the dataset (33.3%), followed by erosion (22.2%), fat (20.7%) and sclerosis (19.3%) (Table 1). While more prevalent in the axSpA positive cohort, lesions typical of axSpA were found in both axSpA-positive and axSpA-negative cases. Complete descriptive data on lesion incidence are available in Tables 1-3.

Regression

Model 1 (Global): In univariable logistic regression of model 1 majority reader agreement variables, subchondral BME, erosion, and fat were all found to be significant [ORs of 8.91 (95% CI 2.56-30.96), 15.78 (95% CI 20.66-120.55), and 14.27 (95% CI 1.86-109.25) respectively for a positive rheumatologist's AxSpA diagnosis.] (Supplementary Table S1). However, when all lesion variables with $p<0.20$ in univariable regression were entered together

into a multivariable model, only subchondral BME was found to be significant (OR 5.10, 95% CI 1.22-21.43) (Supplementary Table S2)

Model 2 (2-Dimensional): Using majority reader agreement data, univariable logistic regression of variables used in model 2 showed statistically significant increased odds of axSpA diagnosis with an increase in number of joint sides or joints with subchondral BME (OR=2.76 95% CI 1.51 to 5.05), inflammation in an erosion cavity (OR=299E+003 95% CI 1.11to 80.7E+009), joint fluid (OR=5.59, 95% CI 1.02 to 30.81), sclerosis (2.43, 95% CI 1.02 to 5.81), erosion (OR=3.27, 95% CI1.39 to 7.71), and fat (OR=2.73, 95% CI=1.23 to 6.08) (Supplementary Table S3). No significant variable emerged when all lesion scores with $p<0.20$ in univariable analysis were entered together into a multivariable model. (Supplementary Table S4)

Model 3 (3-Dimensional): When univariable logistic regression was run on model 3 variables, an increase in subchondral BME (OR=1.90, 95% CI 1.13-3.27), erosion (OR=1.28, 95% CI 1.05-1.56), and fat (OR=1.16, 95% CI 1.00-1.35) was associated with significantly increased odds of axSpA diagnosis. (Supplementary Table S5). Subchondral BME was the only variable associated with an increased odds of axSpA diagnosis in multivariable analysis (OR=2.16, 95% CI 1.10-4.20). (Supplementary Table S6)

CHAID

Using model 1 (global) variables, between 1 and 3 independent variables were retained in the CHAID tree depending on which reader's data was included. Subchondral BME was found to be the most significant predictor of axSpA diagnosis for the initial split of the tree for 4 of 7 readers, as well as the majority reader agreement data. Among these 4 readers and the majority reader agreement data, there was an inconsistency between variables retained in subcategories of cases with and without subchondral BME. After accounting for the presence or absence of BME, further significant predictors included fat (Figure 2), erosion (Figure 5), and bone bud (Figure 9), depending on the reader, while two readers' data showed no further significant predictors (Figures 6,7). The majority reader data model stratified cases in the subchondral BME-absent subcategory by presence of fat, resulting in an OR of 6.38 (95% CI 2.56-30.96) for positive axSpA diagnosis for cases without subchondral BME where fat was present vs. an OR of 0.08 (95% CI 0.02-0.29) for cases with neither subchondral BME nor fat present (Figure 2). Among the 3 readers whose data did not show subchondral BME to be the most significant predictor, erosion was found to be the most significant predictor for 2 readers (Figures 3,8), while fat was found to be the most significant predictor for the other (Figure 4).

CHAID trees produced from model 2 (2D semi-quantitative) variables showed the same structure as global models for 4 of the 7 readers. For two of

the other 3 readers, tree-depth was reduced by one level (Figures 11,12), while for the third, subchondral BME in more than one joint side was required for significant diagnostic prediction. In cases where BME was present in one quadrant or less, inflammation at site of erosion became a significant predictor of axSpA (OR=6.38, Figure 13). The all-reader mean data followed this same pattern, wherein the first tree-split was created by mean subchondral BME score of $\geq$0.43, and inflammation in an erosion cavity >0 was shown to be the most significant predictor of axSpA diagnosis in cases with subchondral BME <0.43 [OR=6.38(2.56-30.96)] (Figure 10).

When detailed semi-quantitative scoring data was entered into the model (3), all but one reader's data showed BME score to be the most significant predictor for axSpA, with no further tree levels (ORs ranging from 9.01-28.54, Figures 15,17-21). The same was true for all-reader mean data [OR=13.75 (3.13-60.35, Figure 14]. A single reader's data produced a tree where a fat score >6 was the most significant predictor of axSpA [OR=14.61(1.91-111.79), Figure 16], with no further predictors detected. Odds ratios for axSpA diagnosis produced from these single-level trees were generally higher than any of the other trees produced by previous models.

Discussion

This study analyzed a multiple-feature axSpA MRI score data set at 3 different levels of scoring granularity using both traditional regression techniques and the Chi-Squared Automatic Interaction Detection (CHAID) algorithm. The objective was to evaluate whether CHAID could play a role in determining the most important features of an image scoring system containing multiple types of lesions potentially associated with a diagnosis. Assuming no a priori knowledge of particular combinations of lesions serving as strong diagnostic predictors compared to single lesion types alone, a comprehensive exploratory regression model would need to include every possible interaction term. The 12-lesion scoring system featured in this paper (models 1 and 2) would require inclusion of 66 interaction terms just to explore all two–variable combinations, or 220 for three-variable combinations, making for near-untenable manual input and difficult interpretation. In contrast, the CHAID package in SPSS 28.0 can search behind the scenes and detect significant interactions of in a matter of a few seconds.

Although these analyses were limited by a sample size of 135, making it less likely for lower levels of the CHAID tree to contain enough cases for any variable to achieve significance, all models produced a tree with at least two levels. The importance of considering predictor combinations is illustrated in both Model 1 (global lesion presence) and Model 2 (number of joints/joint sides affected) analyses of combined reader data, where CHAID produced a tree

depth of two variables with very high contrasting odds ratios at the terminal nodes, while multiple logistic regression found only a single variable significant in model 1 and none in model 2. Given that simple logistic regression found several other statistically significant MRI variables associated with diagnosis, it is reasonable to suspect that with a larger sample size, CHAID may produce more complex trees incorporating these additional variables. Given the relatively large number of lesion types available for analysis, the algorithm makes quick work of checking for all possible combinations of interaction effects when compared to the manual process of creating new variables representing combinations of variable conditions.

Using CHAID to analyze the ASAS MRImagine study data, we were also able to examine how a comprehensive set of MRI lesion variables interact when scored at different levels of granularity (with granularity increasing from model 1 to model 3). When scoring dichotomously for lesion presence on the entire scan (model 1), majority reader data showed that presence of subchondral BME was the strongest predictor of a positive diagnosis, but presence of fat was nearly as predictive of a positive diagnosis in the absence of BME. The number of joint sides with presence of subchondral BME was also shown to be the strongest predictor of axSpA diagnosis in model 2, with presence of erosion in the absence of BME also serving as a strong predictor.

For most readers, as well as mean reader scores, BME again served as a strong predictor in the more granular Model 3 data. When detailed

semi-quantitative scores were entered, CHAID produced single-level trees with cut-offs for BME score greater than 2 for most readers (>=1 for mean scores). Resulting ORs from the nodes of these detailed scoring data trees were generally larger in magnitude than ORs from global trees that included more than one predictor. Considering this, it seems that semi-quantifying lesions, or at least evaluating whether the case meets a certain threshold value for optimal number and/or size of lesions is important for diagnostic discernment. Again, it seems likely that an analysis of a larger data set may produce classification trees with larger combinations of semi-quantitative data and help produce decision-trees with specific cut-off values for several lesion types.

The difference in individual readers' CHAID trees draws attention to possible challenges with reader reliability or scoring styles in this particular methodology. In most cases, predictors selected by majority reader data produced similar diagnostic odds within each individual reader's scoring data. Occasional deviations from the majority reader pattern emerged in readers who were particularly sensitive to certain lesion types (e.g. fat) compared to other readers. Overall, BME emerged as the single most important predictor of axSpA diagnosis in this dataset. In the absence of BME, lesions holding the most predictive power varied amongst readers at all levels of granularity, showing that individual reader style may impact the diagnostic utility of certain lesion types in their particular scoring data. Previous studies on this same data have shown weaker inter-rater reliability in certain lesion types such as backfill and erosion. The fact that these lesion types do not show up in any node of most

readers' classification trees may point to their lower utility in diagnostic discernment or may simply be reflective of their lower occurrence in this particular sample.

Our results generally align with previous analyses of detailed (model 3) scoring data by the ASAS MRI working group (8) that showed BME was more sensitive and specific for axSpA diagnosis at baseline than any other lesion, as well as in other axSpA cohorts[36]. In these previous analyses, erosion was the most specific and sensitive structural lesion for diagnosis when all lesions were considered independently In this study, CHAID identified that fat lesions can suggest a diagnosis of axSpA in patients without current BME, a finding more difficult to uncover by conventional area under the curve or regression analyses of this dataset, though consistent with previous findings in other axSpA cohorts [37] . One major limitation of our current analysis is the exclusion of clinical and other imaging data (eg. radiography of pelvis and spine) from the model—a choice made due to the small sample size available relative to the already large number of MRI variables of interest. The occurrence of MRI lesions in the context of these other clinical manifestations of axSpA must be considered in order to paint a complete picture of their relevance to the diagnostic process ([38–40] Due to potential overfitting of the model and an underpowered sample, analyses should be repeated in a much larger dataset with models containing all relevant non-MRI clinical information, possibly with the aid of future imaging assessment technology (for example, deep-learning

extracted lesion scoring) to ease the burden of expert time required to yield a powerful sample size.

In conclusion, we found that the CHAID algorithm can be helpful in automating analysis of relationships between lesion types and diagnosis within complex multivariable image scoring system data sets across different levels of granularity. Our findings highlight the strength of the relationship of BME and fat lesions on MRI to rheumatologist's diagnosis of SpA.

Acknowledgements

We would like to thank all ASAS MRImagine study central image readers for providing their scoring data for this study.

Declaration of conflict of interest

The authors have no conflict of interest to declare.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Table 1.1. Descriptive table for Model 1: Number (%) of cases where majority readers agree lesion present (global) on SIJ MRI

Lesion	**All (n=135)**	**AxSpA=No (n=38)**	**AxSpA=Yes (n=97)**
Subchondral BME	45 (33.3%)	3 (7.9%)	42 (43.3%)
Inflammation positive MRI according to ASAS definition	25 (18.5%)	0 (0%)	25 (25.8%)
Inflammation in an erosion cavity	15 (11.1%)	0 (0%)	15 (15.5%)
Joint Fluid	12 (8.9%)	2 (5.3%)	10 (10.3%)
Capsule Inflammation	3 (2.2%)	0 (0%)	3 (3.1%)
Enthesitis	5 (3.7%)	0 (0%)	5 (5.2%)
Sclerosis	26 (19.3%)	3 (7.9%)	23 (23.7%)
Erosion	30 (22.2%)	1 (2.6%)	29 (29.9%)
Fat Lesion	28 (20.7%)	1 (2.6%)	27 (27.8%)
Fat Lesion ≥1cm	14 (10.4%)	1 (2.6%)	13 (13.4%)

Fat metaplasia in an erosion cavity (Backfill)	12 (8.9%)	1 (2.6%)	11 (11.3%)
Bone Bud	1 (0.7%)	1 (2.6%)	1 (1.0%)
Ankylosis	6 (4.4%)	1 (2.6%)	6 (6.2%)

Table 1.2. Descriptive table for Model 2: Mean (SD) of All-Reader Mean 2-Dimensional lesion score

Lesion	**All (n=135)**	**AxSpA=No (n=38)**	**AxSpA=Yes (n=97)**
Subchondral Bone Marrow Edema	0.73 (1.05)	0.22 (0.41)	0.94 (1.15)
Inflammation in an erosion cavity	0.15 (0.33)	0.00 (0.02)	0.21 (0.37)
Joint Fluid	0.18 (0.32)	0.09 (0.23)	0.22 (0.34)

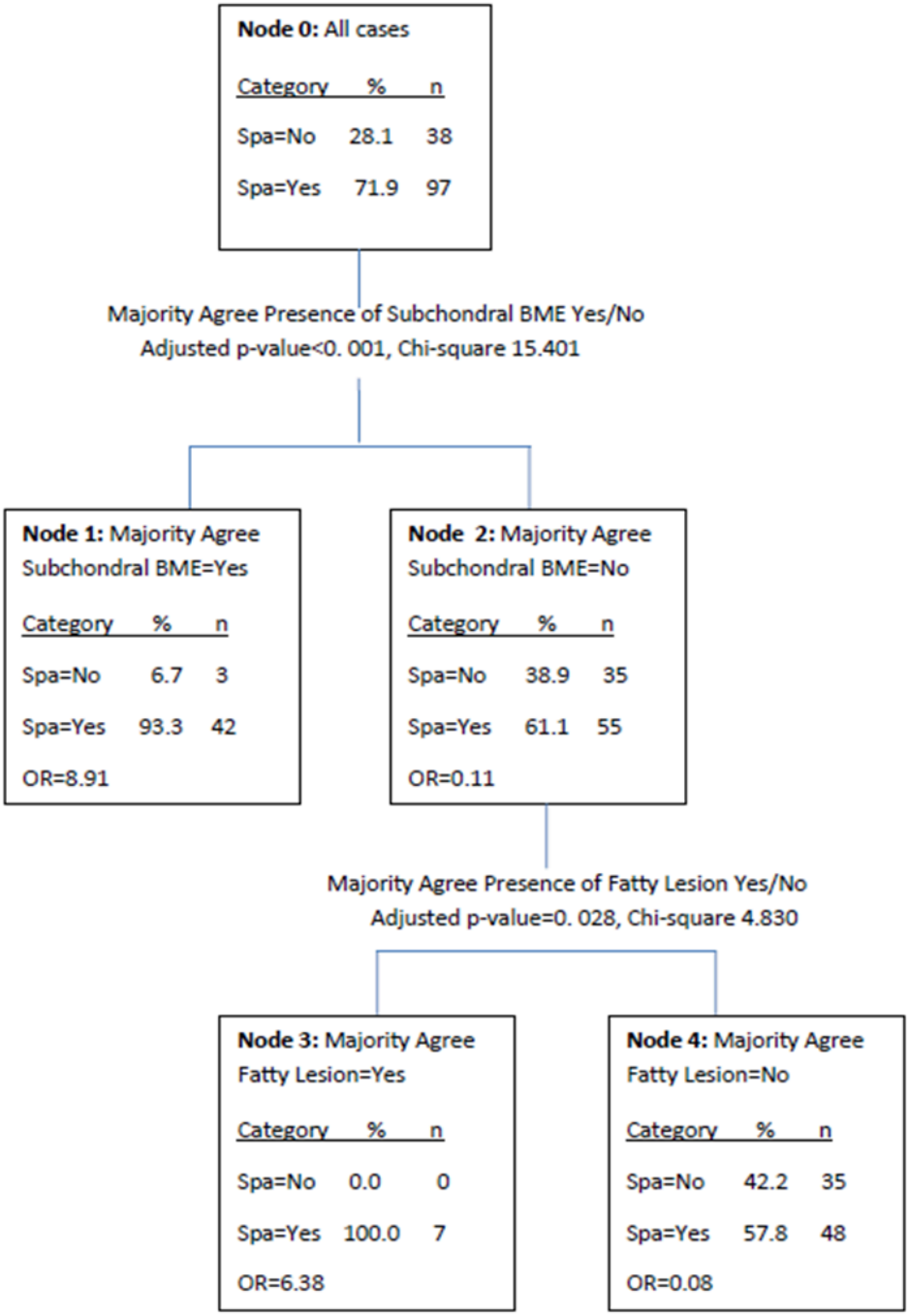
Capsule Inflammation	0.04 (0.18)	0.00 (0.00)	0.05 (0.21)
Sclerosis	0.40 (0.18)	0.23 (0.55)	0.46 (0.60)
Erosion	0.56 (0.91)	0.16 (0.53)	0.72 (0.98)
Fat Lesion	0.50 (0.90)	0.15 (0.64)	0.63 (0.95)
Fat Lesion≥ 1cm	0.23 (0.58)	0.06 (0.28)	0.30 (0.65)
Fat metaplasia in an erosion cavity (Backfill)	0.13 (0.32)	0.03 (0.16)	0.17 (0.36)
Bone Bud	0.05 (0.14)	0.04 (0.12)	0.05 (0.14)
Ankylosis	0.07 (0.28)	0.01 (0.05)	0.09 (0.33)

Table 1.3. Descriptive table for Model 3: Mean (SD) of All-reader Mean 3-Dimensional detailed lesion scores.

Lesion	**All (n=135)**	**AxSpA=No (n=38)**	**AxSpA=Yes (n=97)**
BME	3.97 (10.32)	0.34 (0.62)	5.39 (11.88)
Sclerosis	1.75 (3.91)	0.86 (2.18)	2.11 (4.37)
Erosion	2.46 (4.64)	0.62 (2.29)	3.18 (5.11)
Fat Lesion	2.73 (6.13)	0.77 (4.33)	3.50 (6.56)
Fat Lesion >1cm	1.37 (3.74)	0.45 (2.76)	1.72 (4.01)
Fat metaplasia in an erosion cavity (Backfill)	0.55 (1.59)	0.12 (0.70)	0.72 (1.80)

Bone Bud	0.07 (0.19)	0.06 (0.17)	0.07 (0.20)
Ankylosis	0.51 (3.59)	0.01 (0.07)	0.71 (4.23)

Figure 1.1. Chi-Square Automated Interaction Detection (CHAID) Classification tree for rheumatologist's diagnosis of spondyloarthritis, using majority reader agreement data for global presence of MRI lesions (model 1).

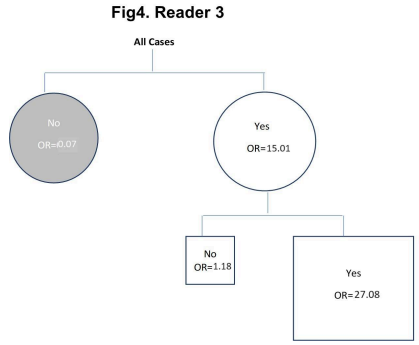


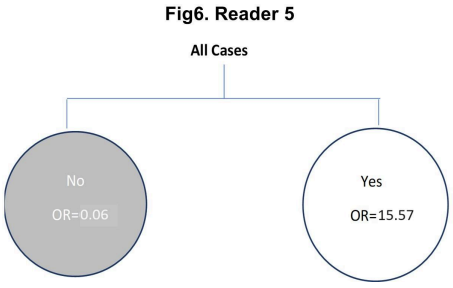
Figures 1.2-1.20: Geometrical representation of variation in Reader Classification Trees

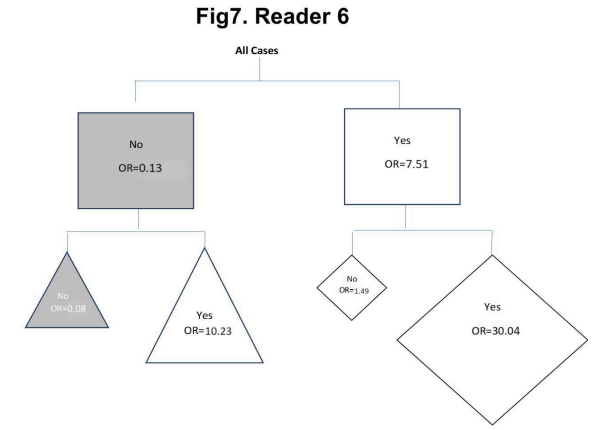
Legend:

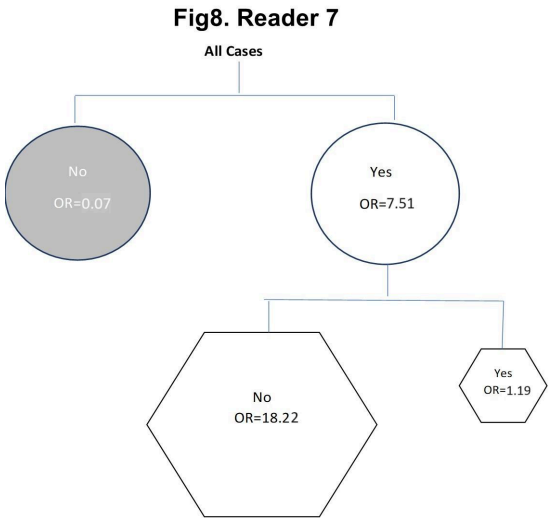
○=Bone Marrow Edema □ =Erosion △ =Fat Lesion ◇ =Sclerosis ⬡ =Bone Bud ✞=Inflammation in an erosion cavity	Shape size is relative to magnitude of Odds Ratio for axSpA diagnosis within each reader's decision tree. White shapes represent OR>1 (positive association with axSpA+ diagnosis) Grey shapes represent OR<1 (negative association with axSpA+ diagnosis)

Figures 1.2-1.9. Model 1: Global Presence of Lesions (Yes/No).









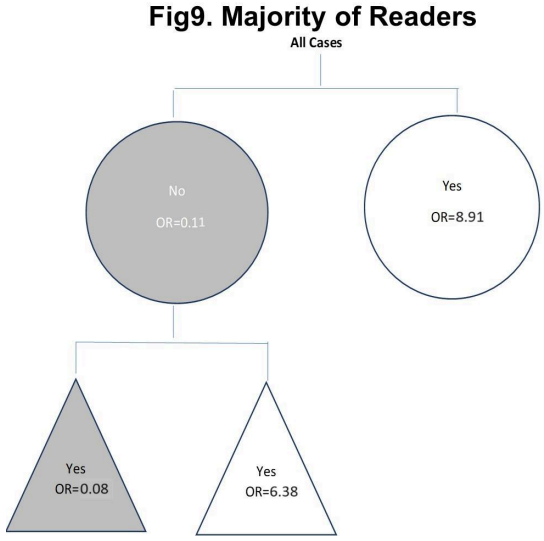


Figure 1.10-1.13. Model 2: 2-Dimensional Quadrant Scores. Reader trees omitted where identical to Model 1.

Fig10. Reader 1

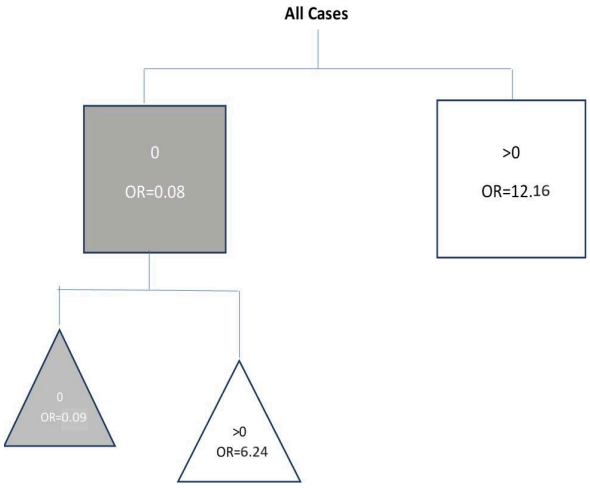


Fig 11.Reader 2

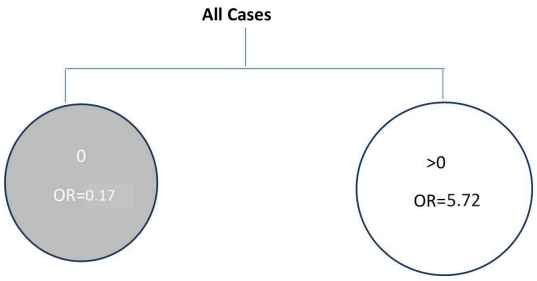


Fig12. Reader 6

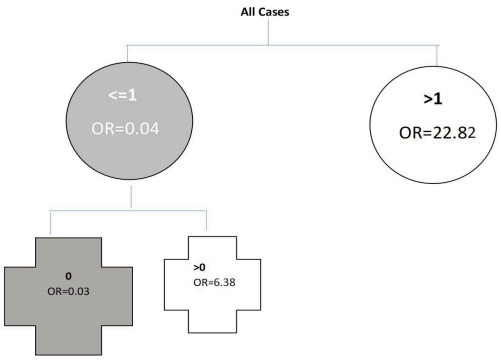
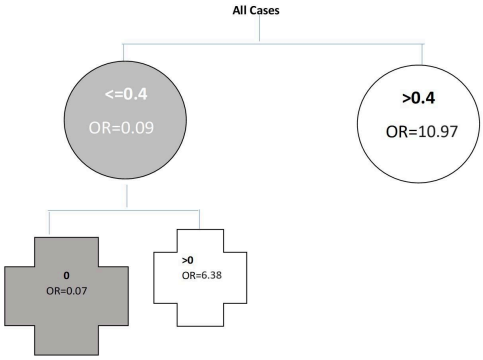


Fig13. Mean of All Readers



Figures 1.14-1.21. Model 3: 3-Dimensional Detailed Quadrant Scoring

Fig14. Reader 1

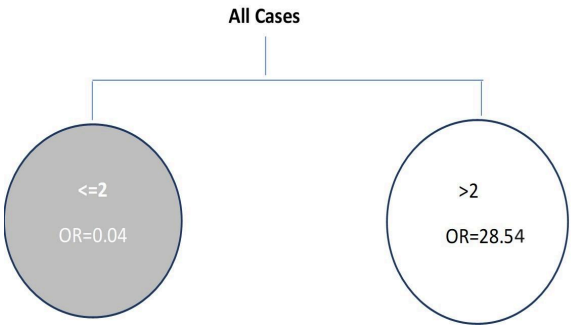


Fig15. Reader 2

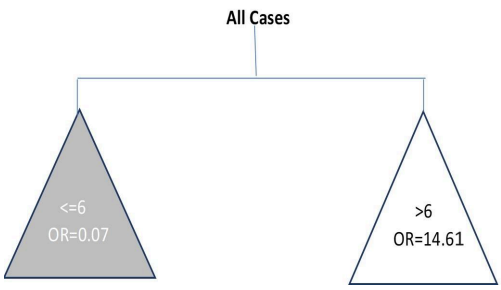
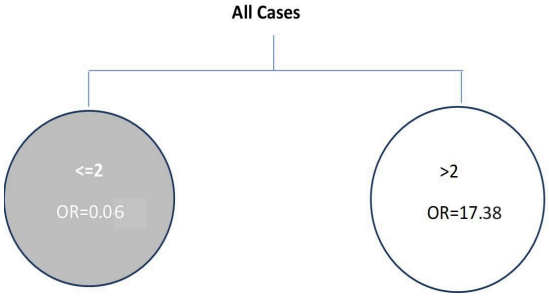


Fig16.Reader 3

Fig17. Reader 4



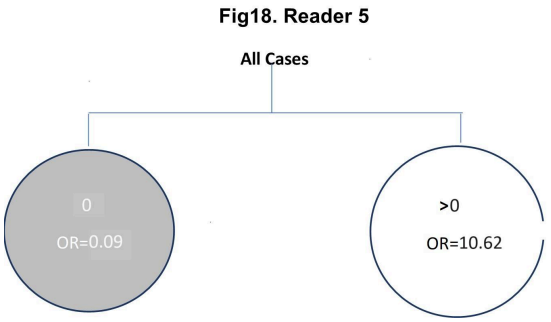


Fig18. Reader 5

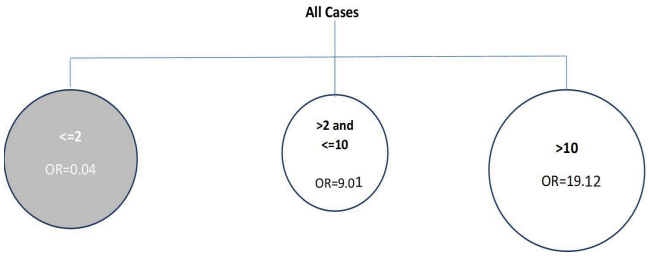


Fig19. Reader 6

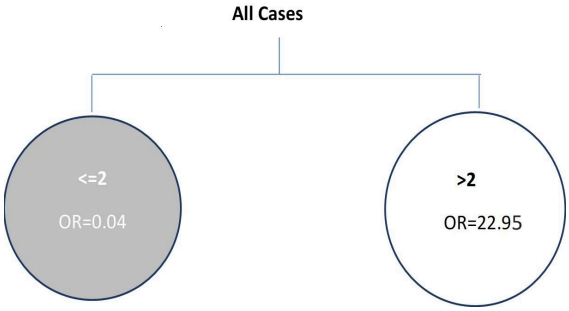


Fig20. Reader 7

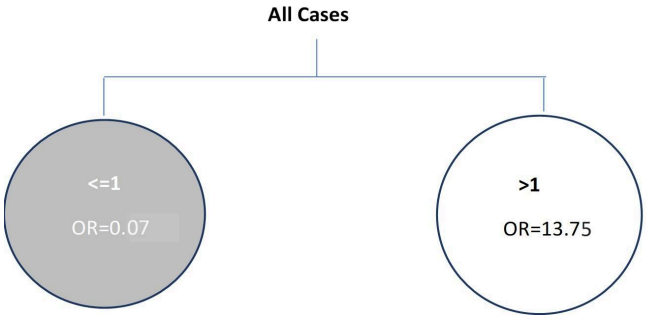


Fig 21. Mean of All Readers

Supplementary Tables

Table 1.S1. Univariate logistic regression analysis of majority reader agreement data on global presence of lesions (Model 1)

Variable (Majority of Readers Agree Lesion is Present)	**Odds Ratio**	**95% CI**	**p-value**

Subchondral Bone Marrow Edema	8.91	2.56 to 30.96	0.001
Inflammation in an erosion cavity **	--	-	-
Joint Fluid	2.79	0.60 to 12.98	0.192
Capsule Inflammation**	--		-
Enthesitis	--	--	-
Sclerosis	3.63	1.02 to 12.89	0.166
Erosion	15.78	2.07 to 120.55	0.047
Fat Lesion	14.27	1.89 to 109.25	0.011
Fat Lesion >=1cm	5.73	0.72 to 45.40	0.099
Backfill	4.73	0.59 to 38.00	0.144
Bone Bud **	--	--	-
Anklylosis**	--	--	-

* variables with p<0.20 entered into multivariable model.

** model could not be fitted due to perfect or quasi-perfect separation in sample

Table 1.S2. Multivariable logistic regression of majority reader global lesion presence data, entering variables with p<0.20 in univariable logistic regression (Model 1)

Variable (Majority of Readers Agree Lesion is Present)	**Odds Ratio**	**95% CI**	**p-value**
Subchondral Bone Marrow Edema	8.91	2.56 to 30.96	0.001
Inflammation in an erosion cavity **	1.41×10^9	--	--
Joint Fluid	2.79	0.60 to 12.98	0.192
Capsule Inflammation**	--	--	--
Enthesitis **	--	--	--
Sclerosis	3.63	1.02 to 12.89	0.166
Erosion	15.78	2.07 to 120.55	0.047
Fat Lesion	14.27	1.89 to 109.25	0.011
Fat Lesion >=1cm	5.73	0.72 to 45.40	0.099
Backfill	4.73	0.59 to 38.00	0.144
Bone Bud **	--		--
Anklylosis **	--		--

** model could not be fitted due to perfect or quasi-perfect separation in sample

Table 1.S3. Univariable Logistic Regression of Mean 2-Dimensional Quadrant Scores (Model 2)

Variable (Mean of 7 Readers)	**Odds Ratio**	**95% CI**	**p-value**
Subchondral Bone Marrow Edema	**3.99**	**1.62 to 9.79**	**0.003***
Inflammation in an erosion cavity	**299E+003**	**1.11 to 80.70E+009**	**0.048***
Joint Fluid	**5.59**	**1.02 to 30.81**	**0.048***
Capsule Inflammation**	--	--	--
Sclerosis	**2.43**	**1.02 to 5.81**	**0.046***
Erosion	**3.27**	**1.39 to 7.71**	**0.007***
Fat Lesion	**2.73**	**1.23 to 6.08**	**0.014***
Fat Lesion >1cm	4.34	0.94 to 20.11	0.061*
Backfill	17.35	0.86 to 348.76	0.062*
Bone Bud	1.65	0.09 to 30.60	0.738
Ankylosis	48.94	0.11 to 21891.33	0.212

* variables with $p<0.20$ entered into multivariable model.

** model could not be fitted due to perfect or quasi-perfect separation in sample

Table 1.S4. Multivariable Logistic Regression of mean 2-Dimensional Quadrant Scores (Model 2)

Variable (7 Reader Mean 2-Dimensional Score)	**Odds ratio**	**95% CI**	p-value
Subchondral Bone Marrow Edema	2.97	0.91 to 9.65	0.070
Inflammation in an erosion cavity	167E+006	0.00 to 8.70E+18	0.133
Joint Fluid	0.49	0.05 to 5.38	0.561
Sclerosis	1.01	0.32 to 3.17	0.984
Erosion	0.23	0.04 to 1.42	0.114
Fat Lesion	1.97	0.46 to 8.35	0.359
Fat Lesion >1cm	0.93	0.08 to 10.19	0.952
Backfill	2.32	0.03 to 176.71	0.704

Table 1.S5. Univariable logistic regression of mean 3-Dimensional semi-quantitative lesion scores (Model 3)

Variable (7-Reader Mean)	**Odds Ratio**	**95% CI**	**p-value**

Subchondral Bone Marrow Edema	**1.92**	**1.13 to 3.27**	**0.015***
Sclerosis	1.16	0.96 to 1.40	0.120*
Erosion	**1.28**	**1.05 to 1.56**	**0.016***
Fat Lesion	**1.16**	**1.00 to 1.35**	**0.047***
Fat Lesion ≥1cm	1.18	0.97 to 1.45	0.101
Backfill	1.76	0.94 to 3.29	0.079*
Bone bud	1.47	0.18 to 12.20	0.723
Ankylosis	2.46	0.13 to 1039.19	0.283

* variables with p<0.20 entered into multivariable model.

Table 1.S6. Multivariable logistic regression of Model 3 7-reader mean semi-quantitative lesion scores

Variable (7 Reader Mean 3D Detailed Score)	**Odds ratio**	**95% CI**	**P**
Subchondral Bone Marrow Edema	**2.16**	**1.09 to 4.28**	**0.027**
Sclerosis	0.86	0.67 to 1.10	0.229
Erosion	0.89	0.66 to 1.19	0.415
Fat Lesion	1.53	0.70 to 3.32	0.285
Fat Lesion >1cm	0.58	0.20 to 1.66	0.307
Backfill	1.25	0.54 to 2.87	0.600

* variables with p<0.20 entered into multivariable model.

CHAPTER 2

Title: Reliability and sources of discrepancy in granular semi-quantitative bone marrow lesion scoring on knee MRI: a Study using KIMRISS

Background:

Subchondral bone marrow lesions (BML) on fluid-sensitive MRI are an important manifestation of osteoarthritis (OA) [41,42]) and have been shown to be associated with pain [43,44],[45], structural changes [46,47], and eventual knee replacement surgery [48]. BMLs may be present to varying extents and in varying patterns throughout the knee and can have myriad causes, both traumatic and non-traumatic [41]. Their presence as it relates to OA diagnosis, prognosis, and disease severity has been described qualitatively [41], and several different semi-quantitative scoring methods have been developed to measure BMLs in OA, including the Whole Organ MRI Score (WORMS) [49], Boston-Leeds Osteoarthritis Knee Score (BLOKS) [50]), the MRI Osteoarthritis Knee Score (MOAKS) [23], and the Outcome Measures in Rheumatology (OMERACT) Knee Inflammation MRI Scoring System (KIMRISS) [24,51]. All of these scoring methods semi-quantify BML to different degrees of detail using pre-specified rules for

regional subdivision and, in some cases, recording lesion size within subregions.

Data derived through these types of semi-quantitative medical image scoring systems can be an important tool in hypothesis-driven or exploratory analysis of factors associated with diagnosis, disease severity, disease progression, and treatment efficacy [52], and the assignment of a semi-quantitative image score using pre-established rules has the potential to elucidate in detail the difference in lesion severity between cases and/or longitudinal changes in lesions. An operationally defined numerical score allows for a moresomewhat objective and nuanced analysis of imaging findings as they relate to laboratory markers, other imaging modality findings (eg. the relationship of active manifestations of disease on MR to accumulated structural changes visible on MRI or plain radiographs [46,53]), or patient self-report measures (eg. pain, function, and quality of life).

Furthermore, scoring methodology-defined subdivisions of an anatomical region may bring to light relationships of lesion presence at a particular location to certain outcomes not captured by a score summarizing the broader region. It has already been demonstrated, for example, that BML in the medial compartment of the knee is more significantly associated with eventual knee replacement surgery compared to other regions [54].

Meaningful analysis of semi-quantitative data requires demonstrable consistency and reliability in scoring between different readers. As long as

human expert readers are required to generate this data, the value of time spent scoring images is a serious consideration and demands optimization. Small differences between cases or time points may be detected, and the significance of lesion presence in particular quantities or locations within an anatomical structure may be discovered or clarified through detailed scoring data analyses, but only if there is reasonable assurance that this data is reliable. If the data is not reliable enough to be used in such analyses, the semi-quantitative assessment may not provide enough value for time spent.

With all of this in mind, the design of any semi-quantitative scoring methodology must strike a balance between reliability of scoring data output, utility of the data generated, and feasibility for raters. Ideally, a scoring system provides enough information to the end-user that its value in research and analysis exceeds that of any dichotomous or qualitative assessment that could be extracted from a standard radiologist's report.

This paper aims first and foremost to evaluate the reliability aspect of BML scoring optimization by examining data derived from the KIMRISS methodology. KIMRISS provides data on an increased number of predefined scoring regions in the sagittal plane over previous methodologies, resulting in smaller scoring regions that are each assessed dichotomously for lesion presence rather than both presence and size. Previous studies have shown very good inter-rater reliability for the total KIMRISS score summarizing the extent of BML across all regions and slices, but analyses of the reliability for smaller subregions have yet to be published. Here, we assess reliability of scoring

down to the smallest of sub-regions to explore the extent to which a more granular analysis of relationships between lesion location and outcomes is feasible. Further to the reliability assessment, we have examined some potential sources of discrepancy in scoring, as these could be a guiding factor in modification or exclusion of certain scoring regions from such in-depth analyses.

Methods:

KIMRISS Scoring:

8 readers trained and calibrated in the KIMRISS methodology scored BML on n=61 knee MRIs from the osteoarthritis initiative (OAI) dataset. Using a web-based module, readers place prefabricated grid overlays onto the patella, femur, and tibia and indicate the slice numbers defining the borders of the lateral, intercondylar, and medial regions of the tibiofemoral joint (anchor slices). Scoring was carried out in each grid square on all slices between the outer lateral and outer medial boundaries.

Data Analysis:

Inter-rater reliability of scoring by all readers was assessed via intraclass correlation coefficient (ICC) for the entire knee; each component of the knee (patella, femur, tibia); the medial, lateral and intercondylar portions of the whole knee and its components; the all-slice sum of each of the 28 individual grid squares defined by the KIMRISS overlay (Figure 1); individual

tibial and femoral grid square scores in the medial, lateral, and intercondylar compartments; and finally the aforementioned medial and lateral compartment scores further subdivided into 6 groups of slices medially to laterally (Figure 2).

In search of possible sources of inter-rater discrepancy due to slice selection, we used a subset of 3 readers to tally up instances of disagreement on the slice assignment for medial, lateral, and intercondylar grid anchors and the absolute difference in slice number between readers.

Additionally, to check for possible discrepancy in granular scoring due to differing placement and/or sizing of the grid within a single sagittal slice, we attempted to identify instances where the same lesion may have been assigned to different grid squares by different readers. In order for an instance to be counted, we used individual slice data to locate grid squares scored positively by one reader and negatively by a second reader, where an immediately adjacent grid square was scored positively by the second reader but not the first reader. (Adjacent square definition and syntax shown in supplementary figure). A percentage of possible adjacent grid square scores for the same lesion between readers was calculated in each region by dividing the number of instances of adjacent grid square scoring between readers by the total number of scoring discrepancies between those readers in that region.

Results:

The 8 reader two-way mixed absolute agreement ICC (95% CI) for the total sum of BML across all regions was 0.72 (0.58 to 0.82). The ICC (95% CI)

for the total sum of BML was highest in the femur [0.84 (0.76-0.90)] compared to the tibia and patella [ICC (95%) 0.73 (0.63-0.81), and 0.80 (0.71-0.87), respectively].

Scoring ranges and ICCs of individual grid squares across all slices are shown in Table 1 and Figure 1, respectively. The 8-reader ICC for BML at each scoring location was very good (>=0.80) or good (0.70-0.80) for 17/28 (61%) of the grid regions . The lowest ICCs occurred at the inferior patella [P4; ICC=0.52], a small interior region of the femur [mean (SD) ICC=0.49 (0.26)], and the two most posterior regions of the tibia [mean (SD) ICC=0.41 (0.26) and 0.31 (0.33)]. Lower reliability was also found in regions of the tibia not immediately adjacent to the tibiofemoral joint (ICC range: 0.60-0.67) as well as posterior non-articular regions of the femur (ICC range: 0.59-0.65). Regions with a broader range of overall incidence of BML tended to demonstrate higher reliability, but this was not always the case. For example, the data showed similar mean (SD) scores in the weight bearing regions of the femur vs. the below subchondral regions of the tibia, but much higher reliability in the former.

After further subdividing regions medially to laterally, highest overall reliability amongst all KIMRISS grid squares occurred in the central slices of the lateral compartment of the femur (median ICC=0.72) and the innermost slices of the medial compartment of the tibia (median ICC=0.42). (Table 2.2).

Lowest overall reliability occurred in the central slices of the medial compartment of the femur (median ICC=0.15) and the medialmost slices of the medial compartment of the tibia (median ICC= 0.28) (full schematics of each coronal subregion shown in supplementary figures 1 a-f).

An analysis of anchor placement showed the greatest discrepancy in slice assignment at the medial edge of the intercondylar region in the femur, with a slice discrepancy of >3 slices in up to 56% of cases. The readers' placement of all other anchors (lateral edge, lateral intercondyle, medial intercondyle in both femur and tibia and the medial and lateral edges in the patella) fell within 3 slices of one another 92-100% of the time.

Taking into account the findings for all reader pairs within each grid square, adjacent grid square analysis showed that placement of lesions in neighbouring regions between readers occurred at a mean rate of less than 4.3% of scoring discrepancies at each location. However, certain reader pairs may have placed the same lesion into discrepant regions at a rate of more than 10% of total scoring discrepancies at T5, T11, FP3, and FT2 (17.9%, 16.7%, 12.5%, and 11.6%, respectively).

Discussion:

Inter-rater reliability of granular scoring locations varied substantially between different regions of the knee.

Reduced intra-class correlation coefficients for BML were seen as the regions analyzed became smaller. This is somewhat expected, as the numerator of the ICC formula includes between-case variation which has a tendency to decrease as the maximum possible score decreases. [55] Smaller regions with a narrower range of possible scores require lower between-reader variability (entered into the denominator of the equation) to produce high ICCs.

That being said, the reliability of certain regions remained more robust even when subdivided into smaller coronal sections compared to other regions with similar range and between-case variability of scoring. This was particularly true of the anterior portion of the femur in the lateral compartment. The reliability of some of the smaller regions was incalculable due to limited incidence of BML scoring within those locations, especially at the medialmost slices of the femur and posterior tibia.

The anchor slice selection discrepancy at the intercondylar edge of the medial condyle may be a cause of discrepancy within the medial compartment, although reliability suffered more in the outermost medial compartment compared to the inner medial compartment. This may be due to the distribution of scores within these regions.

Scoring discrepancy due to variation in grid sizing or placement, identified by positive scores in adjacent grid squares by different readers, did not appear with high frequency, but did happen to occur among certain reader pairs in regions that had lower reliability at all levels of analysis. This may

provide a case for combining data at these regions, such as T5/T55 at the posterior tibia prior to including it in any location-specific outcomes analyses.

Achievement of fair to good all-slice score reliability in 25/28 KIMRISS grid regions suggests this scoring data carries a reasonable degree of validity. Further division of this data into the coronal plane presents certain reliability challenges particularly in the medial slices, which may be a function of the distribution of BMLs within this particular dataset. As previous studies show the importance of medially located BMLs in prediction of disease progression, it may be prudent to optimize scoring subregions within this compartment for reliability so more detailed analyses can be carried out.

As advancements are made in automated full quantification of BML volume in the knee[56], it may be important to clarify if and how quantification should be segmented into regions that provide the clearest picture of the image's relevance to clinical outcomes. Semi-quantitative scoring datasets such as the one studied here can provide insight into how this should be done.

Figure 2.1. KIMRISS grid region labels with 8-reader mean ICC from baseline knee MRIs from the Osteoarthritis Initiative dataset (n=61).

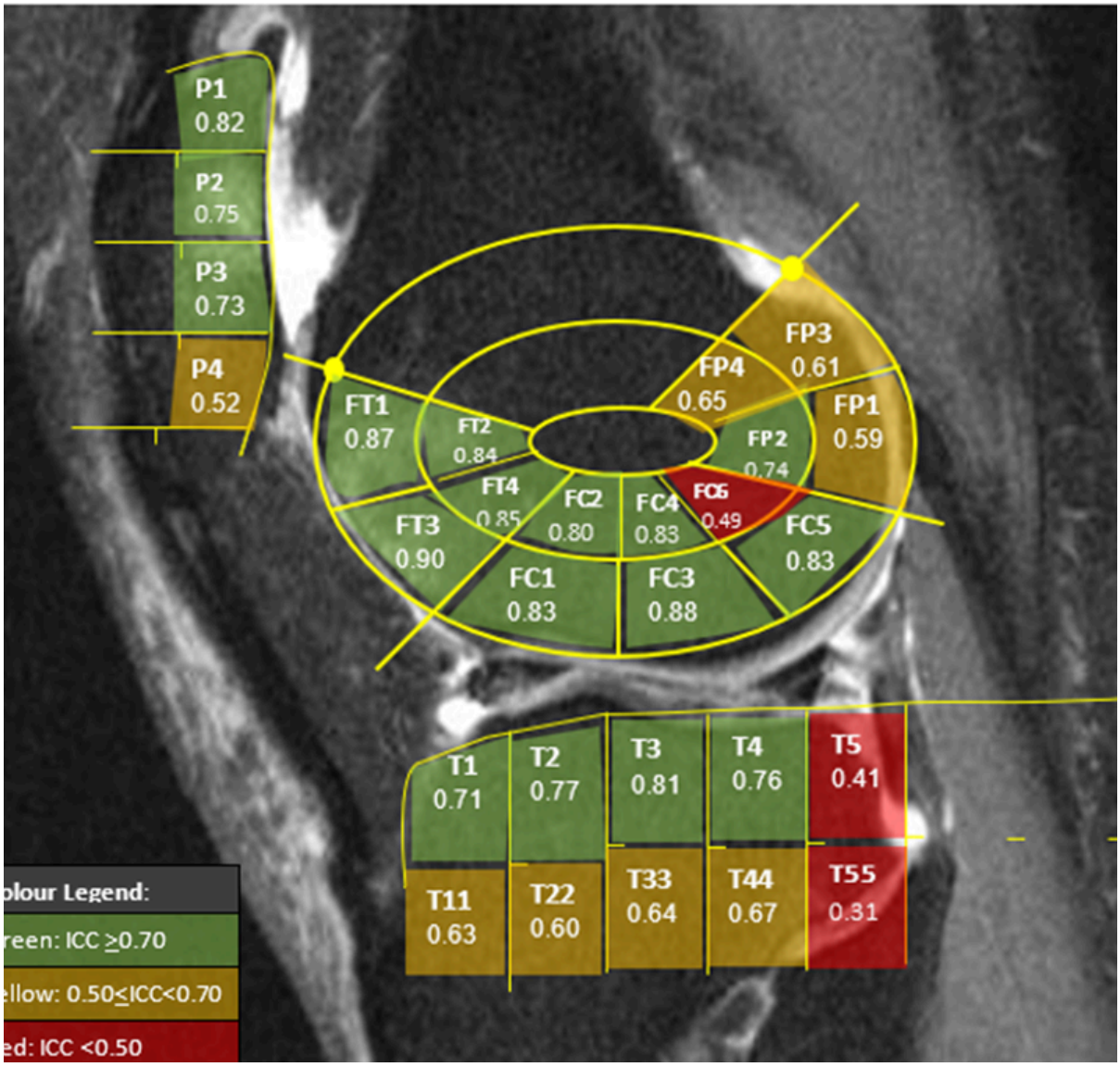


Table 2.1. 8-reader mean (SD) BML score at each KIMRISS grid region (colour coded according to ICC in Fig 1.)

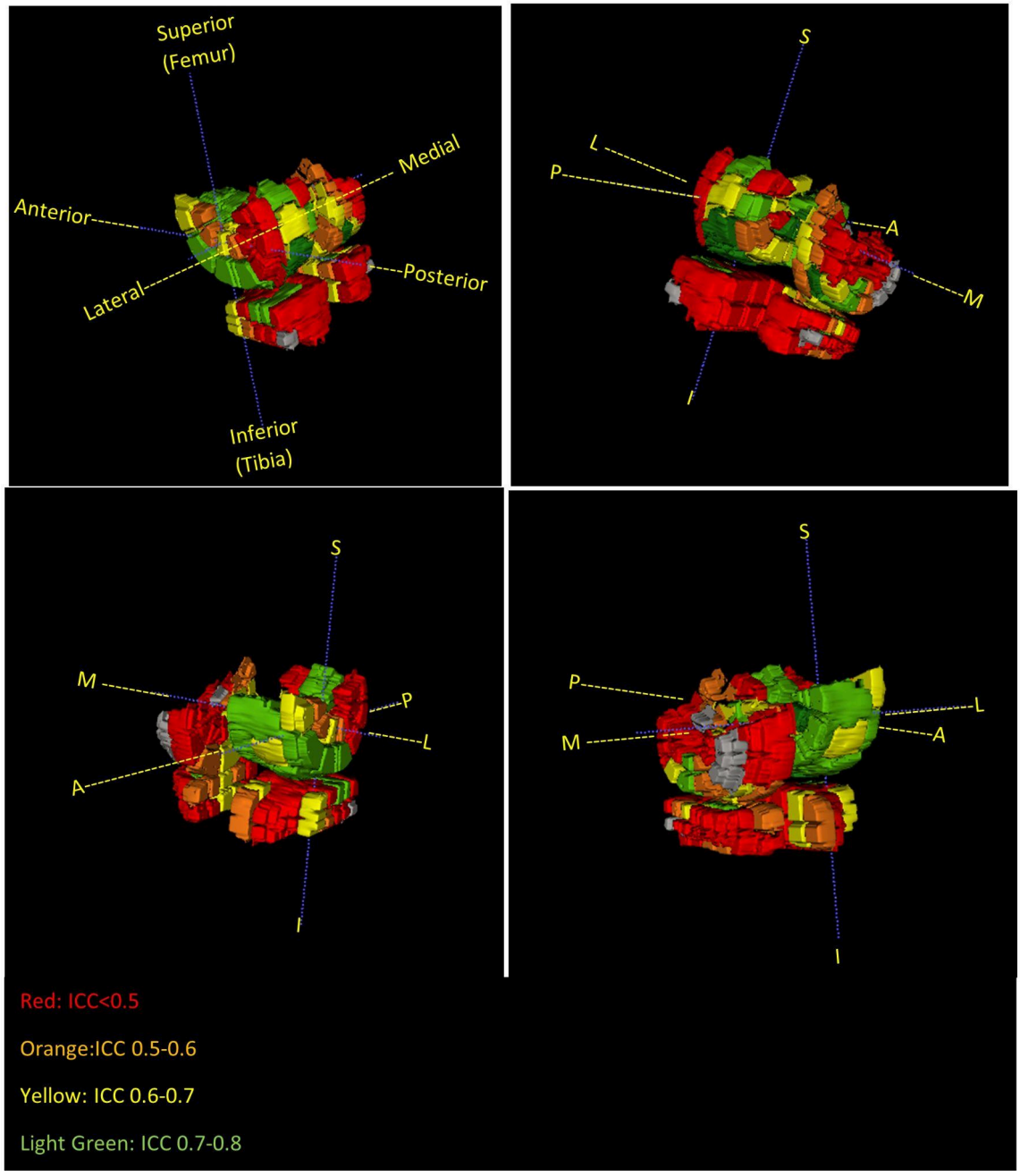
Patella		Femur			Tibia		
Region	Mean (SD)	Region		Mean (SD)	Region		Mean (SD)
P1	2.0 (2.5)	Trochlear	FT1	1.8 (2.7)	Subchondral Region	T1	1.6 (2.8)
P2	2.5 (2.6)		FT2	0.6 (2.0)		T2	1.8 (3.0)
P3	1.9 (2.3)		FT3	2.6 (3.0)		T3	1.9 (3.1)
P4	0.6 (1.1)		FT4	0.8 (2.1)		T4	1.4 (2.5)
		Weight-bearing	FC1	1.3 (2.2)		T5	0.5 (1.0)
			FC2	0.6 (1.9)	Below Subchondral Region	T11	0.6 (1.7)
			FC3	0.8 (2.0)		T22	0.8 (1.9)
			FC4	0.3 (1.4)		T33	0.8 (1.8)

			FC5	0.5 (1.6)		T44	0.6 (1.5)
			FC6	0.2 (0.8)		T55	0.2 (0.6)
		Posterior	FP1	0.4 (0.9)			
			FP2	0.2 (0.7)			
			FP3	0.2 (0.6)			
			FP4	0.1 (0.5)			

Table 2.2. Summary of ICCs across individual femoral and tibial KIMRISS grid subregions, subdivided medially to laterally.

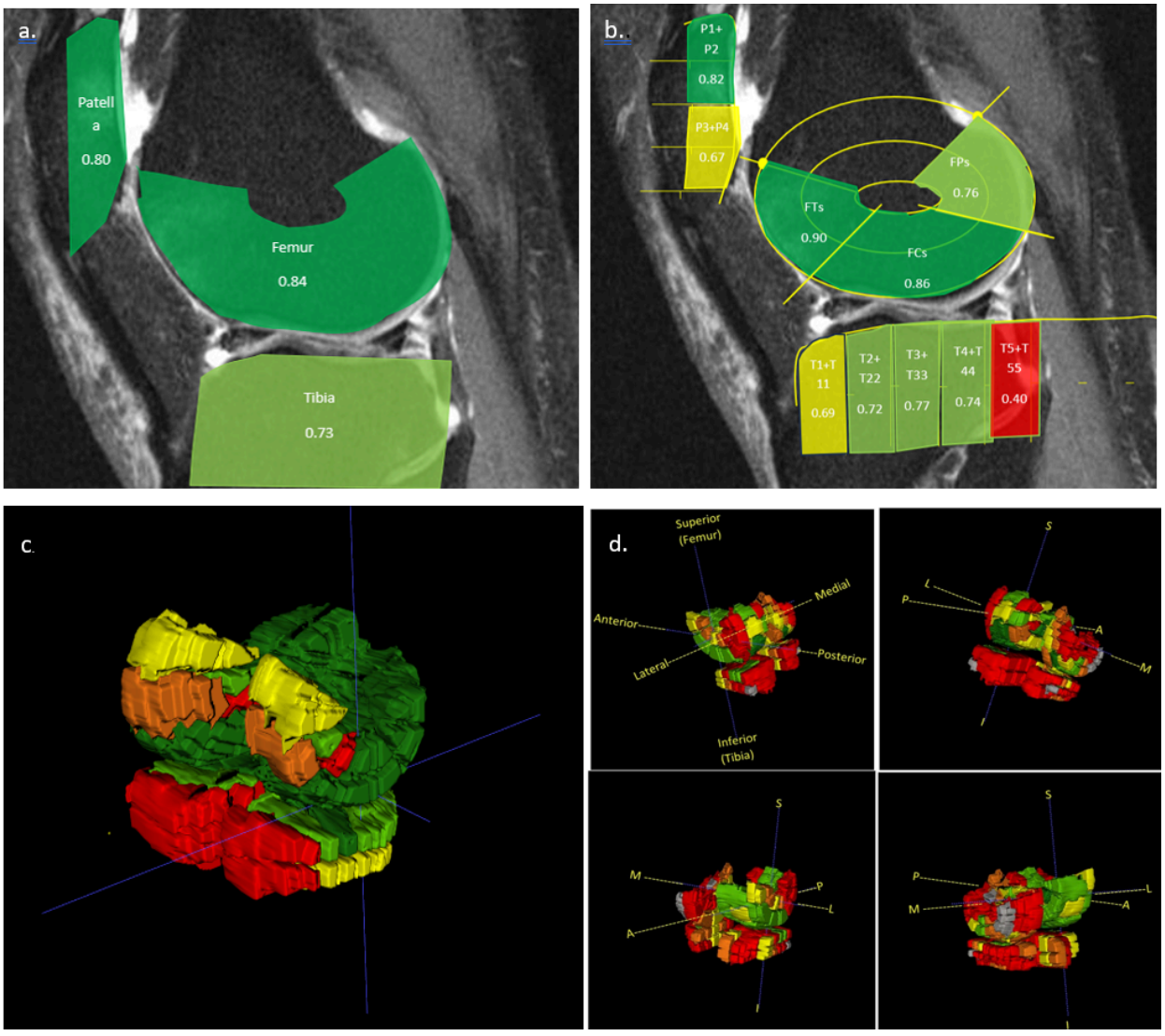
		All Lateral Slices	Lateral Lateral (3 most lateral slices)	Lateral mid	Lateral Inner (3 slices closest to intercondylar region)	All Medial Slices	Medial Inner (3 slices closest to intercondylar region)	Medial mid	Medial Medial (3 most medial slices)
8-Reader ICC in individual KIMRISS Femoral subregions (n=14)	Median	0.69	0.55	0.72	0.68	0.33	0.42	0.15	0.38
	IQR	0.55 to 0.77	0.10 to 0.68	0.69 to 0.80	0.55 to 0.79	0.060 to 0.59	0.23 to 0.54	0.098 to 0.60	0.00 to 0.60
	Min	0	0	0.56	0.07	-0.01	0	-0.01	0
	Max	0.94	0.75	0.94	0.92	0.84	0.69	0.84	0.76
8-Reader ICC in individual KIMRISS Tibial subregions (n=10)	Median	0.44	0.54	0.35	0.46	0.42	0.54	0.46	0.28
	IQR	0.26 to 0.56	0.29 to 0.69	0.20 to 0.47	0.26 to 0.53	0.27 to 0.58	0.34 to 0.60	0.28 to 0.59	0.16 to 0.49
	Min	0	0.1	0	0	0	0.24	0.05	0
	Max	0.75	0.75	0.7	0.62	0.76	0.67	0.76	0.65

Figure 2.2. 3D views of of reliability in KIMRISS grid squares subdivided medially to laterally



Supplementary Figure 2.S1 a-d. 8-Reader intraclass correlation coefficients of bone marrow lesion scoring in the knee at different levels of granularity. a.

all-slice whole bone scores; b. all-slice scores divided into large subregions; c. all-slice KIMRISS-defined grid square regions; d. KIMRISS-defined grid square regions subdivided into slice groupings in the coronal plane a. b. c. d.



Supplementary Figure 2.S2. Methodology used to assess possible assignment of same lesion to different grid squares by different readers.

Square	Adjacent Squares	Formula Template for each reader
FT1	FT2 and FT3	=IF(AND(R1_FT1=1,R2_FT1=0,OR(AND(R2_FT2=1,R1_FT2=0),AND(R2_FT3=1,R1_FT3=0))),1,0)
FT2	FT1 and FT4	=IF(AND(R1_FT2=1,R2_FT2=0,OR(AND(R2_FT1=1,R1_FT1=0),AND(R2_FT4=1,R1_FT4=0))),1,0)
FT3	FT1, FT4, FC1	=IF(AND(R1_FT3=1,R2_FT3=0,OR(AND(R2_FT1=1,R1_FT1=0),AND(R2_FT4=1,R1_FT4=0),AND(R2_FC1=1,R1_FC1=0))),1,0)
FT4	FT2, FT3, FC2	=IF(AND(R1_FT4=1,R2_FT4=0,OR(AND(R2_FT2=1,R1_FT2=0),AND(R2_FT3=1,R1_FT3=0),AND(R2_FC2=1,R1_FC2=0))),1,0)
FC1	FT3, FC2, FC3	=IF(AND(R1_FC1=1,R2_FC1=0,OR(AND(R2_FT3=1,R1_FT3=0),AND(R2_FC2=1,R1_FC2=0),AND(R2_FC3=1,R1_FC3=0))),1,0)
FC2	FT4, FC1, FC4	=IF(AND(R1_FC2=1,R2_FC2=0,OR(AND(R2_FT4=1,R1_FT4=0),AND(R2_FC1=1,R1_FC1=0),AND(R2_FC4=1,R1_FC4=0))),1,0)
FC3	FC1, FC4, FC5	=IF(AND(R1_FC3=1,R2_FC3=0,OR(AND(R2_FC1=1,R1_FC1=0),AND(R2_FC4=1,R1_FC4=0),AND(R2_FC5=1,R1_FC5=0))),1,0)
FC4	FC2, FC3, FC6	=IF(AND(R1_FC4=1,R2_FC4=0,OR(AND(R2_FC2=1,R1_FC2=0),AND(R2_FC3=1,R1_FC3=0),AND(R2_FC6=1,R1_FC6=0))),1,0)
FC5	FC3, FC6, FP1	=IF(AND(R1_FC5=1,R2_FC5=0,OR(AND(R2_FC3=1,R1_FC3=0),AND(R2_FC6=1,R1_FC6=0),AND(R2_FP1=1,R1_FP1=0))),1,0)
FC6	FC4, FC5, FP2	=IF(AND(R1_FC6=1,R2_FC6=0,OR(AND(R2_FC4=1,R1_FC4=0),AND(R2_FC5=1,R1_FC5=0),AND(R2_FP2=1,R1_FP2=0))),1,0)
FP1	FC5, FP2, FP3	=IF(AND(R1_FP1=1,R2_FP1=0,OR(AND(R2_FC5=1,R1_FC5=0),AND(R2_FP2=1,R1_FP2=0),AND(R2_FP3=1,R1_FP3=0))),1,0)
FP2	FC6, FP1, FP4	=IF(AND(R1_FP2=1,R2_FP2=0,OR(AND(R2_FC6=1,R1_FC6=0),AND(R2_FP1=1,R1_FP1=0),AND(R2_FP4=1,R1_FP4=0))),1,0)
FP3	FP1 and FP4	=IF(AND(R1_Fp3=1,R2_FP3=0,OR(AND(R2_FP1=1,R1_FP1=0),AND(R2_FP4=1,R1_FP4=0))),1,0)
FP4	FP2 and FP3	=IF(AND(R1_Fp4=1,R2_FP4=0,OR(AND(R2_FP2=1,R1_FP2=0),AND(R2_FP3=1,R1_FP3=0))),1,0)
T1	T11 and T2	=IF(AND(R1_T1=1,R2_T1=0,OR(AND(R2_T11=1,R1_T11=0),AND(R2_T2=1,R1_T2=0))),1,0)
T11	T1 and T22	=IF(AND(R1_T11=1,R2_T11=0,OR(AND(R2_T1=1,R1_T1=0),AND(R2_T22=1,R1_T22=0))),1,0)
T2	T1, T3, T22	=IF(AND(R1_T2=1,R2_T2=0,OR(AND(R2_T1=1,R1_T1=0),AND(R2_T3=1,R1_T3=0),AND(R2_T22=1,R1_T22=0))),1,0)
T22	T11, T33, T2	=IF(AND(R1_T22=1,R2_T22=0,OR(AND(R2_T11=1,R1_T11=0),AND(R2_T33=1,R1_T33=0),AND(R2_T2=1,R1_T2=0))),1,0)
T3	T2, T4, T33	=IF(AND(R1_T3=1,R2_T3=0,OR(AND(R2_T2=1,R1_T2=0),AND(R2_T4=1,R1_T4=0),AND(R2_T33=1,R1_T33=0))),1,0)
T33	T3, T22, T44	=IF(AND(R1_T33=1,R2_T33=0,OR(AND(R2_T3=1,R1_T3=0),AND(R2_T22=1,R1_T22=0),AND(R2_T44=1,R1_T44=0))),1,0)
T4	T3, T5, T44	=IF(AND(R1_T4=1,R2_T4=0,OR(AND(R2_T3=1,R1_T3=0),AND(R2_T5=1,R1_T5=0),AND(R2_T44=1,R1_T44=0))),1,0)
T44	T4,T33,T55	=IF(AND(R1_T44=1,R2_T44=0,OR(AND(R2_T4=1,R1_T4=0),AND(R2_T33=1,R1_T33=0),AND(R2_T55=1,R1_T55=0))),1,0)
T5	T4 and T55	=IF(AND(R1_T5=1,R2T5=0,OR(AND(R2T4=1,R1_T4=0),AND(R2T55=1,R1_T55=0))),1,0)
T55	T5 and T44	=IF(AND(R1_T55=1,R2T55=0,OR(AND(R2T5=1,R1_T5=0),AND(R2T44=1,R1_T44=0))),1,0)
P1	P2	=IF(AND(R1P1=1,R2P1=0,R2P2=1,R1P2=0),1,0)
P2	P1 and P3	=IF(AND(R1P2=1,R2P2=0,OR(AND(R2P1=1,R1P1=0),AND(R2P3=1,R1P3=0))),1,0)
P3	P2 and P4	=IF(AND(R1P3=1,R2P3=0,OR(AND(R2P2=1,R1P2=0),AND(R2P4=1,R1P4=0))),1,0)
P4	P3	=IF(AND(R1P4=1,R2P4=0,R2P3=1,R1P3=0),1,0)

In KIMRISS, BML is scored on on T2-Fat Suppressed or T1 (check) MRI in the sagittal plane. Users place grid squares

CHAPTER 3

Title: Evaluation of an Artificial Intelligence Algorithm to Generate Semi-quantitative Bone Marrow Lesion Scores

Purpose

Bone marrow lesions (BML) are an important feature on fluid-sensitive magnetic resonance imaging (MRI) of Knee Osteoarthritis (OA) [57], and the size and specific location of BML within the knee may have differential impact on clinical measures and outcomes [43 58,59]. Granular scoring of BML is time-consuming for human readers, but allows for detection of small changes over time in addition to producing a record of precise lesion location. As automated BML scoring methods are developed, detailed lesion location information could be preserved while eliminating the burden of scoring time. Here, we aim to demonstrate the value of recording the precise location of BML compared to less granular approaches, and perform preliminary validation on an AI algorithm which produces semi-quantitative BML scores analogous to those generated by human experts.

Methods

Human reader scoring data analysis

8 calibrated readers scored 62 sagittal knee MRIs from the Osteoarthritis Initiative dataset using the Knee Inflammation MRI Scoring System

(KIMRISS[24,25]), which incorporates an overlay template to score 28 predefined locations (grid squares) in the femur, tibia, and patella for the presence of BML. Mean BML scores were calculated at the following levels of granularity (Figure 1): 1a. whole femur, patella, and tibia across all slices; 1b.femur, patella, and tibia divided into lateral, intercondylar, and medial coronal subregions; 2a.Neighboring grid square scores combined into larger regions across all slices; 2b. combined regions divided into coronal subregions; 3a. A total score for each of the 28 individual grid squares summed across all slices; and 3b. Scores for each individual grid square subdivided into coronal subregions.

Univariable logistic regression for eventual arthroplasty was performed on BML scores at each described level of granularity. To examine the relationship of BML location to patient self-report measures of pain, stiffness, and function, correlations to the relevant Western Ontario and McMaster Universities Arthritis Index (WOMAC)[60] scores were measured at all levels of granularity via Kendall's Tau.

To identify possible interactions between granular BML scoring locations, individual coronally subdivided grid square scores were entered into the Chi-Squared Automatic Interaction Detection (CHAID) algorithm using SPSS v.29.01[31]. to produce a decision-tree for eventual arthroplasty. The minimum number of cases for the parent and child nodes of the CHAID tree were set to 10 and 5, respectively, and 10-fold cross-validation was performed.

Automated BML Scoring evaluation:

A custom iMaskRCNN deep learning model was trained on 700 MRI slices using corresponding pixel-wise bone marrow lesion segmentation derived from expert KIMRISS scores and their corresponding grid template placements.[28] The algorithm was trained to automatically segment bone, cartilage and BMLs, and identify areas of interest (femoral and tibial heads, patella) for automatic placement of the KIMRISS grid template. The combined result of BML segmentation and grid template placement is then translatable to an automated KIMRISS score (iKIMRISS).

This algorithm was first run on the above n=62 osteoarthritis cases with corresponding human reader KIMRISS scoring available. The program generated an output quantifying the total number of pixels contained within the boundaries of each square of the KIMRISS template, along with its prediction of the number of BML pixels within that square. Because the size of each grid square in the template varies between cases and across slices within an individual MRI, a "BML percentage" for each square was calculated by dividing the number of BML pixels by the total number of pixels assigned. This BML percentage was converted to a dichotomous iKIMRISS score for each grid location (0 or 1 for negative and positive BML score) using three different thresholds (>0%, ≥10%, and ≥20%).

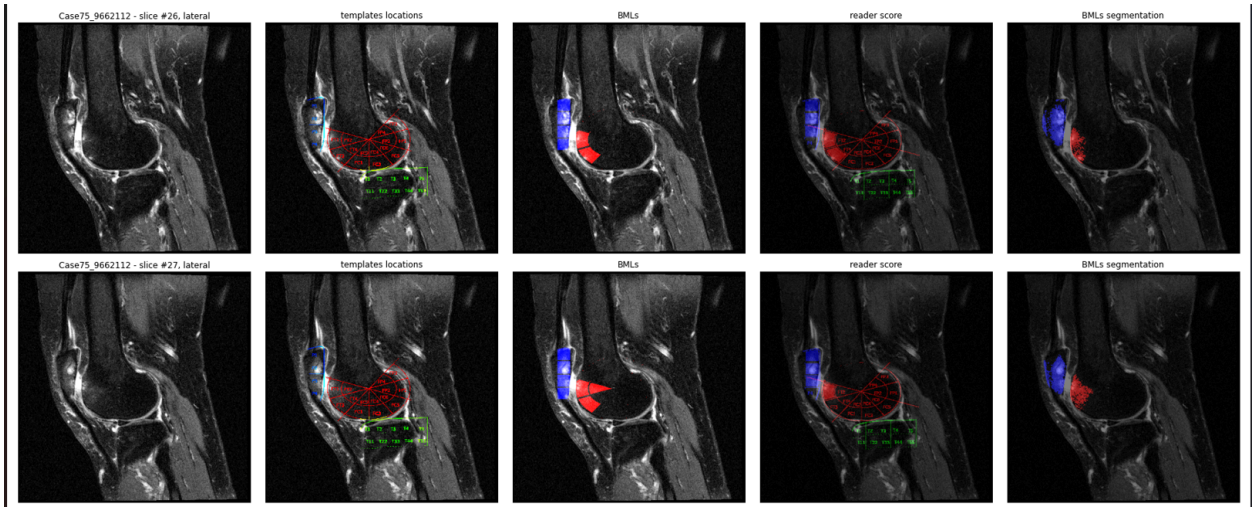
Prior to any formal analysis, the AI-generated data was checked for instances of false negative and false positive scores on the 3922 individual MRI

slices comprising these cases, using majority (≥5) human reader agreement for a positive lesion as the standard for a positive score, and positive lesion assignment from ≤1 reader as the standard for a negative score . False negative and positive scores by the AI were checked for “adjacent grid square assignment” (eg. a false positive score with a false negative score in a neighboring region or vice versa), to determine if errors resulted from discrepant grid template placement rather than discrepancy in lesion detection.

Images were checked visually to identify any other underlying causes of disagreements, and the AI developer used this feedback to adjust the algorithm to improve accuracy. The results displayed in this manuscript are derived from the third and latest adjustment to the algorithm (Figure)

Figure 3.1. Bone marrow lesions, extracted by deep-learning in lateral femur and patella, with automatically placed KIMRISS grid overlay and accompanying human reader score assignment. Slice 26 (top row) demonstrates full agreement with expert human reader data, while slice 27 (bottom) shows partial agreement. Note that the “BML” schematic in the third column represents positive grid scores assigned by deep-learning according to the presence of any amount of BML within the boundaries of each region, prior to

setting pixel area percentage thresholds for positive BML score.



Intra-class correlation coefficients [61]for reliability compared to 8-reader mean human KIMRISS scores were calculated at the level of whole knee, whole bone, grouped neighboring grid square regions, and individual grid squares across all MRI slices, as well as in lateral, intercondylar, and medial compartments. Analyses were repeated on scores generated from different BML pixel percentage thresholds to determine which threshold produced the most reliable KIMRISS translation.

Analysis of large dataset:

The deep-learning iKIMRISS algorithm was run on baseline MRI from n=1631 knees from the Osteoarthritis Initiative dataset where accompanying clinical and follow-up data on arthroplasty were available. Patients younger than 50 years of age were excluded due to their low likelihood of qualifying for arthroplasty due to age.

A rough evaluation of consistency with human reader BML assessment was performed using a subset of n=836 subjects with human reader-generated MRI Osteoarthritis Knee (MOAKS) BML scores available in the dataset [23]. Kendall's tau rank correlations were calculated for iKIMRISS scores for the entire femur, tibia, and patella vs. MOAKS readers' evaluation of the number of bone marrow lesions in each of these regions. Correlation calculations were also performed on the lateral and medial compartments of each bone , and the number of lesions in the supraspinous tibia (MOAKS) was compared to the iKIMRISS score in the intercondylar tibia. Analogous MOAKS data on the intercondylar region of the femur was not available for analysis. More granular analysis of smaller regions within each bone was not possible due to differences in scoring region assignment between the two methodologies.

To evaluate predictive utility of the deep learning-extracted iKIMRISS BML scores for eventual knee replacement, logistic regression was performed at all levels of scoring location granularity for the outcome of any knee replacement surgery recorded within the ten year follow-up period. Associations with WOMAC pain, stiffness, and function scores were evaluated using Kendall's Tau rank correlation.

Finally, the CHAID algorithm with 10-fold cross-validation was applied to coronally subdivided individual grid square iKIMRISS BML scores to assess whether lesion presence in any particular combination of locations is predictive of eventual arthroplasty. In light of the increased sample size compared to the

previous human reader data analysis, minimum case numbers for parent and child nodes were increased to 200 and 100, respectively.

Results:

Human Reader Scoring Data:

In the small test dataset, 13 (21%) patients underwent subsequent arthroplasty.

Logistic regression performed on whole bone and combined region all-slice scores (1a. and 2a. described above) found no significant relationships between BML and subsequent arthroplasty. From the all-slice scores for individual grid squares (3a. above), one anterior tibial region, T2, was significant [Odds Ratio(OR)(95% CI): 1.2 (1.0-1.5)]. (Figure 3.3)

When the same analyses were performed on coronally subdivided scores , no lateral or intercondylar regions emerged as significant predictors). Within medial slices, whole bone scores (1b. above) had no significant relationship to arthroplasty, while two combined grid regions (2b. above) in the anterior tibia were significantly associated with arthroplasty [T1/T11: OR (95% CI): 1.5 (1.09-2.2) and T2/T22: OR (95% CI):1.3 (1.0-1.7)]. Stronger associations with arthroplasty were found in the medial portion of individual grid square scores (3b. above), namely in one small weight-bearing region of the femur [FC1: OR (95% CI): 2.6 (1.2-5.5)], and one small region of the posterior femur [FP3: OR (95% CI) 3.8 (1.3-10.1)] (Figure 3.4).

Figure 3.2. Map of BML scoring regions analyzed at three different levels of granularity

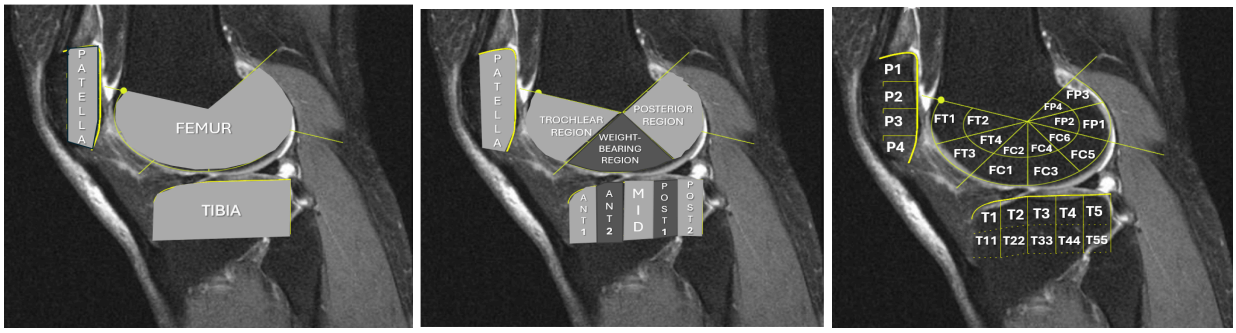


Figure 3.3. Significant Odds Ratios for arthroplasty from bone marrow lesion across all scoring regions (n.s.=not significant)

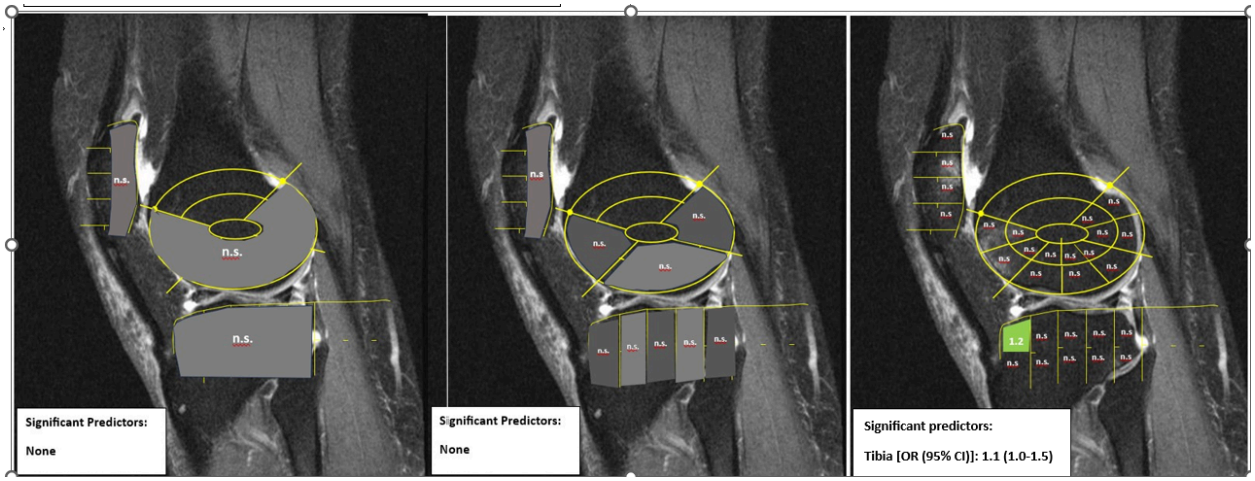
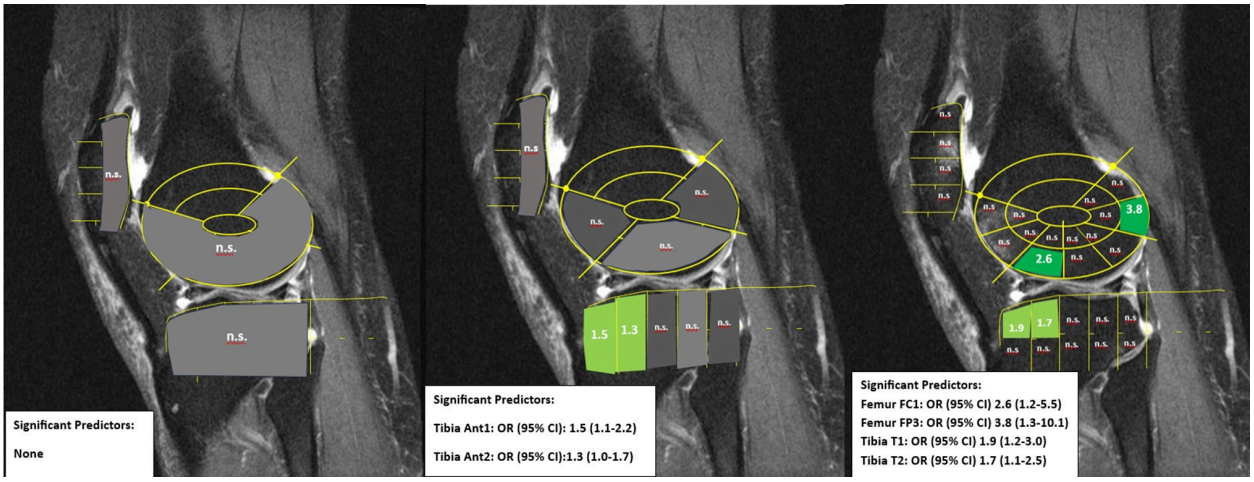


Figure 3.4. Significant Odds Ratios for arthroplasty from bone marrow lesion scores across coronally-subdivided regions. Results shown are from medial slices only, as analysis of lateral and intercondylar scores yielded no significant predictors.

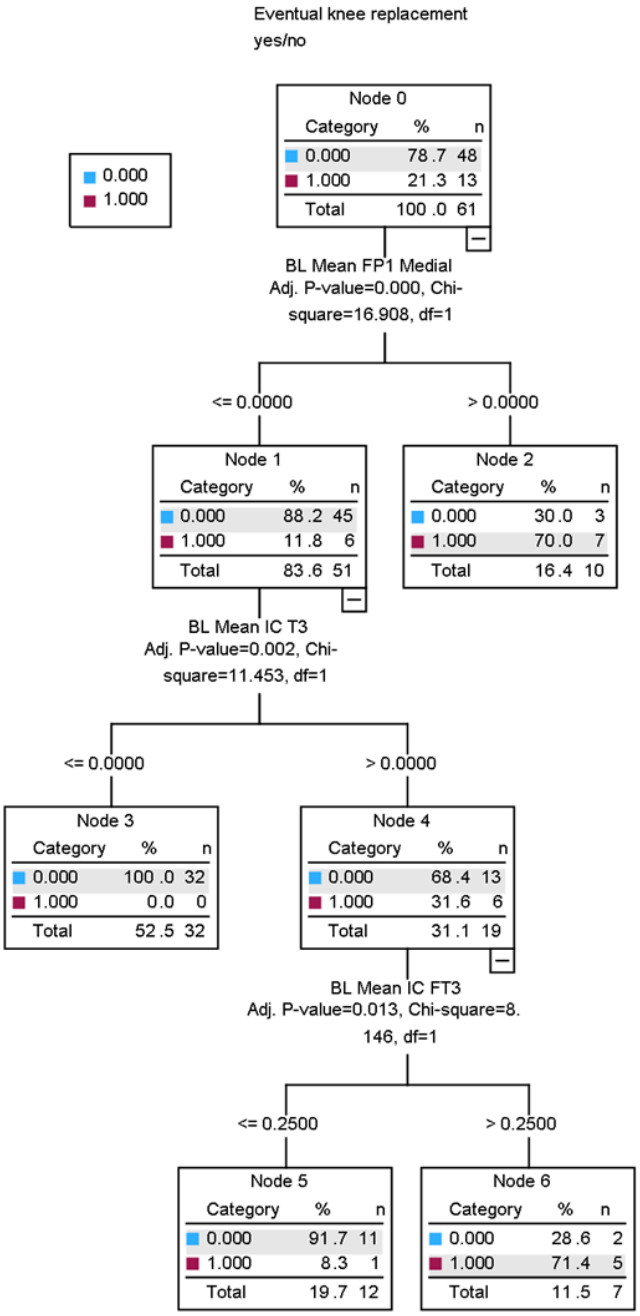


The overall strength of of Kendall's Tau correlation for BML to WOMAC Pain increased as smaller regions were analyzed (Table 2), with the strongest correlation occurring in the medial weight-bearing femur [FC5: 0.53 (95% CI 0.38-0.74)]. The same pattern was observed in correlations to WOMAC Stiffness and Function, with the same medial femoral region bearing the highest significance [FC5: 0.45 (95% CI 0.38-0.74) and 0.53 (95% CI 0.41-0.63) for stiffness and function, respectively].

The CHAID Decision Tree Model showed that, when examined in combination, scores at the medial posterior femur (FP1), intercondylar central tibia (T3), and trochlear region of the femur (FT3) are the best predictors of eventual arthroplasty, producing an area under the curve (AUC) of 0.91 (95%

CI 0.81-0.97), with sensitivity (95% CI) of 0.92 (0.74-0.99) and specificity (95% CI) of 0.90 (0.77-0.97). (Figure 3.5)

Figure 3.5. CHAID-generated decision tree for eventual knee replacement considering coronally-subdivided individual KIMRISS grid square scores derived from 8-reader mean human scoring data.



Deep Learning-Generated iKIMRISS data:

Reliability in comparison to human data:

Descriptive data for lesion incidence among both human and algorithm-generated scoring is available in **Table 3.1**. In the n=62 cases with complete human-generated KIMRISS data available,the algorithm achieved poor reliability when compared to total human KIMRISS scores at the level of the whole joint, requiring a BML pixel-labeling threshold of >=20% to for positive score assignment in each square in order to reach a maximum ICC (95% CI) of 0.25 (0-0.47). When analyzing each of the femur, tibia, and patella on its own, promising moderate agreement with human scoring was found in the tibia, producing an ICC (95% CI) of 0.61 (0.43-0.75) when using a 20% BML pixel labeling threshold for each grid square. Scoring was least reliable in the femur [all slice ICC (95%) of 0.17 (0-0.40) with 20% BML pixel labeling threshold], and suboptimal in the patella [all slice ICC (95%) of 0.31 with 20% BML pixel labeling threshold]. When whole bone scores were subdivided into lateral, medial, and intercondylar compartments, highest agreement with human scoring was found in the medial femur and tibia [ICC (95%): 0.41 (0.18-0.6) and 0.65 (0.48-0.78), respectively]. Reliability in the patella was highest in the lateral compartment [ICC (95%): 0.5 (0.28-0.66)].

After further reducing the region size into groups of neighbouring grid squares, the trochlear (FT) and posterior (FP) regions of the femur showed good reliability in the lateral slices only [ICC 95%: 0.78 (0.66-0.86) and 0.74 (0.59-0.84), respectively], while the central weight-bearing portion of the femur (FCs) only achieved an ICC (95%) of 0.43 at best in the medial compartment. iKIMRISS scores in the tibia showed good reliability in the two most anterior regions (T1T11 and T2T22), achieving ICCs as high as 0.79 (0.67-0.87) in T2T22 when considering all slice scores, and 0.71 in lateral T1T11 when coronally subdivided. By comparison, ICCs were reduced in the central and posterior regions of the tibia, ranging from 0.11-0.64 for regions T33T33,T4T44, and T5T55.

At the level of the individual grid squares, highest reliability was again seen in the tibia, specifically the anterior portions of the bone directly adjacent to the joint space (T1, T2), with all-slice ICCs (95% CI) of 0.78 (0.66-0.87), and 081 (0.70-0.88), respectively. Higher reliability was seen in the medial and intercondylar slices than the lateral slices. Very poor reliability was shown in each of the central rid regions (FPs), with ICCs ranging from 0.01-0.19 when considering all slices, reaching only as high as 0.34 (95% CI 0.09-0.55) in the lateral portion of FC1. Scores in the posterior region showed similarly poor ICCs, ranging from 0.13-0.31 when considering all slices, but moderate improvement when considering lateral slices only, reaching as high as 0.56 (95% CI 0.35-0.71) and 0.74 (95% CI 0.6-0.84) for FP2 and FP3, respectively. One trochlear region, FT1, showed moderate reliability in all-slice scores {ICC

(95%) 0.68 (0.52-0.80)] and very good reliability in lateral slices only [0.89 (0.82-0.93)], but ICCs elsewhere in the trochlea were otherwise low (0.00-0.37) when considering either all-slice or coronally subdivided scores. There was moderate agreement in the upper half of the patella [95% CI 0.52 (0.3-0.68) and 0.69 (0.53-0.81) for all slice scores in P1 and P2, respectively], with reduced agreement in P3 and P4 [0.47 (0.25-0.65) and 0.35 (0.1-0.56), respectively]. Medial slices in these regions had very poor reliability (ranging from 0.0-0.25) when compared to lateral slices (ranging 0.42-0.61).

Table 3.1. Descriptive data of distribution of human KIMRISS scores and deep learning-extracted iKIMRISS scores for n=61 test cases at the level of whole bone and individual grid squares.

	Mean Human Reader KIMRISS score								iKIMRISS score							
	All Slices		Lateral		Medial		Intercondylar		All Slices		Lateral		Medial		Intercondylar	
Region	Range	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)
Whole Femur	0-99.5	10.4 (14.8)	0-82.5	6.6 (12.1)	0-37.5	2.2 (6.1)	0-17	1.5 (2.9)	0-87	28.6 (20.5)	0-54	13.9 (12.5)	0-81	11.5 (15.1)	0-21	3.2 (4.2)
Whole Tibia	0-67.8	8.1 (13.8)	0-30	2.1 (5.3)	0-39.4	3.3 (8.4)	0-22	2.9 (5.2)	0-122	18.7 (24.3)	0-51	4.8 (8)	0-71	8 (14.9)	0-46	5.9 (9)
Whole Patella	0-75	12.2 (13.9)	0-21.6	4.3 (5.6)	0-1.5	0.1 (0.3)	0-11.1	1.5 (2.2)	0-60	13.4 (11.9)	0-60	9.2 (11)	0-7	0.8 (1.6)	0-17	3.4 (4.7)
FC1	0-12.3	1.3 (2.2)	0-10.1	0.7 (1.7)	0-4.5	0.3 (0.9)	0-3.6	0.3 (0.7)	0-10	1.1 (1.8)	0-3	0.4 (0.7)	0-9	0.7 (1.7)	0-0	0 (0)

FC2	0-13.5	0.6 (1.9)	0-10.1	0.4 (1.4)	0-2.4	0.1 (0.4)	0-3.4	0.1 (0.5)	0-7	0.8 (1.4)	0-4	0.3 (0.7)	0-7	0.5 (1.2)	0-0	0 (0)
FC3	0-10.4	0.8 (2)	0-6.9	0.3 (1.1)	0-1 0.4	0.5 (1.7)	0-0	0 (0)	0-7	0.4 (1)	0-1	0 (0.1)	0-7	0.3 (1)	0-0	0 (0)
FC4	0-9.4	0.3 (1.4)	0-9.4	0.2 (1.2)	0-5.5	0.1 (0.7)	0-0	0 (0)	0-8	1.1 (2)	0-8	0.7 (1.6)	0-6	0.4 (1.3)	0-2	0 (0)
FC5	0-9.6	0.5 (1.6)	0-6	0.2 (0.8)	0-9.6	0.4 (1.4)	0-0	0 (0)	0-20	2.6 (3.7)	0-9	1.4 (2.4)	0-7	0.9 (1.8)	0-6	0 (0)
FC6	0-4.4	0.2 (0.8)	0-3.3	0.1 (0.4)	0-4.4	0.1 (0.7)	0-0	0 (0)	0-18	5.9 (4.3)	0-12	2.7 (2.9)	0-8	1.6 (2.3)	0-6	0 (0)
FP1	0-4.6	0.4 (0.9)	0-4.6	0.1 (0.6)	0-3.5	0.2 (0.7)	0-0	0 (0)	0-9	1.9 (2)	0-5	1.1 (1.1)	0-9	0.9 (1.8)	0-0	0 (0)
FP2	0-5.1	0.2 (0.7)	0-5.1	0.1 (0.7)	0-1.9	0.1 (0.3)	0-0	0 (0)	0-9	1.8 (1.7)	0-4	0.6 (0.8)	0-8	1.2 (1.6)	0-0	0 (0)
FP3	0-4.3	0.2 (0.6)	0-4.3	0.1 (0.5)	0-1.6	0.1 (0.3)	0-0	0 (0)	0-9	1.6 (1.6)	0-7	0.4 (1)	0-9	1.2 (1.4)	0-0	0 (0)
FP4	0-3.3	0.1 (0.5)	0-3.3	0.1 (0.4)	0-1.1	0 (0.1)	0-0	0 (0)	0-3	0.6 (0.8)	0-1	0.1 (0.3)	0-3	0.4 (0.8)	0-0	0 (0)
FT1	0-10.6	1.8 (2.6)	0-9.5	1.5 (2.4)	0-1.3	0.1 (0.2)	0-0	0 (0)	0-13	2.2 (3)	0-9	1.7 (2.5)	0-6	0.4 (1.1)	0-5	0 (0)
FT2	0-13.8	0.6 (2)	0-10.1	0.5 (1.5)	0-0.6	0 (0.1)	0-3.6	0.1 (0.5)	0-22	5.2 (4.8)	0-9	2.8 (3.1)	0-10	1.5 (2.2)	0-12	0.9 (2.1)
FT3	0-12.9	2.6 (3)	0-10.3	1.8 (2.5)	0-1.4	0.1 (0.3)	0-5.8	0.7 (1.1)	0-13	3.6 (3.3)	0-9	2.1 (2.4)	0-7	1.2 (1.9)	0-7	0.3 (1.2)
FT4	0-14.4	0.8 (2.1)	0-10.3	0.6 (1.6)	0-1.1	0 (0.2)	0-4.1	0.2 (0.6)	0-10	1.4 (1.9)	0-3	0.6 (0.7)	0-9	0.9 (1.8)	0-0	0 (0)
P1	0-9.9	1.6 (2.4)	0-8.5	1.2 (1.9)	0-0.6	0 (0.1)	0-4.1	0.4 (0.8)	0-18	2.2 (3.4)	0-18	1.6 (3)	0-2	0.1 (0.4)	0-5	0.5 (1.2)
P2	0-10.6	3.2 (3.5)	0-5.9	1.5 (2)	0-0.8	0 (0.1)	0-5.5	0.6 (1)	0-18	4.7 (4.3)	0-18	3.5 (3.9)	0-3	0.2 (0.6)	0-5	1 (1.6)

P3	0-8.5	1.5 (2.2)	0-6.5	1.3 (1.8)	0-0.4	0 (0.1)	0-2.4	0.4 (0.6)	0-14	5.5 (3.6)	0-14	3.5 (3.5)	0-3	0.4 (0.8)	0-6	1.5 (1.8)
P4	0-8.1	1.9 (2.6)	0-3	0.3 (0.7)	0-0.3	0 (0)	0-2.1	0.2 (0.4)	0-10	1.7 (2.3)	0-10	1.2 (2)	0-1	0.1 (0.2)	0-4	0.5 (0.9)
T1	0-12.4	1.5 (2.7)	0-5.9	0.2 (0.8)	0-5.8	0.7 (1.3)	0-6.4	0.6 (1.3)	0-12	2.4 (3.6)	0-3	0.3 (0.6)	0-7	1.3 (2)	0-6	0
T11	0-11.3	0.5 (1.6)	0-5.6	0.1 (0.7)	0-2.4	0.2 (0.5)	0-5.4	0.3 (0.8)	0-10	1.3 (2.5)	0-4	0.2 (0.7)	0-5	0.5 (1.2)	0-5	0
T2	0-12.6	1.6 (2.9)	0-5.1	0.3 (0.9)	0-4.6	0.5 (1.1)	0-4.6	0.5 (1.1)	0-14	3 (3.5)	0-5	0.7 (1.2)	0-9	1.8 (2.4)	0-5	0.6 (1.3)
T22	0-10.3	0.7 (1.8)	0-5.8	0.2 (0.8)	0-4.5	0.2 (0.8)	0-4.5	0.2 (0.8)	0-13	1.7 (3)	0-8	0.5 (1.3)	0-8	0.8 (1.8)	0-5	0.4 (1.2)
T3	0-10.8	1.4 (2.8)	0-6.5	0.4 (1.1)	0-7.5	0.5 (1.6)	0-3.8	0.5 (0.9)	0-15	3 (3.7)	0-5	0.9 (1.5)	0-11	1.2 (2.3)	0-6	0.9 (1.4)
T33	0-7.5	0.6 (1.6)	0-2.4	0.2 (0.5)	0-6.4	0.3 (1)	0-3.1	0.2 (0.5)	0-14	1.9 (3.2)	0-8	0.7 (1.6)	0-8	0.8 (1.8)	0-5	0.4 (1.1)
T4	0-8.8	1 (2.1)	0-7.6	0.3 (1.1)	0-6.9	0.4 (1.2)	0-3	0.4 (0.7)	0-16	2.4 (3.3)	0-9	0.8 (1.5)	0-9	0.8 (1.9)	0-6	0.8 (1.3)
T44	0-7.6	0.4 (1.2)	0-1.5	0.1 (0.3)	0-4.5	0.2 (0.8)	0-2.6	0.2 (0.5)	0-15	1.8 (2.9)	0-8	0.3 (1.2)	0-7	0.6 (1.6)	0-6	0.8 (1.4)
T5	0-4.3	0.4 (0.9)	0-3.1	0.1 (0.5)	0-3.6	0.2 (0.6)	0-1.5	0.1 (0.3)	0-17	0.8 (2.5)	0-10	0.4 (1.5)	0-7	0.4 (1.2)	0-2	0.1 (0.4)
T55	0-3.5	0.1 (0.5)	0-0.1	0 (0)	0-1.6	0.1 (0.2)	0-1.8	0.1 (0.3)	0-12	1.3 (2.3)	0-6	0.3 (1)	0-7	0.4 (1.3)	0-6	0.6 (1.3)

Table 3.2. 2-way consistency intra-class correlation coefficients (ICC) for deep learning-generated **whole bone iKIMRISS scores** compared to mean of 8 human reader scores on n=62 cases with complete human scoring data available. ICCs are shown for iKIMRISS scores calculated according to different

thresholds for the percentage of BML pixels in each grid square required to assign a positive score (20%, 10%, 0%).

	20% BML pixel threshold		10% BML pixel threshold		0% BML pixel threshold	
	ICC	**95% CI**	**ICC**	**95% CI**	**ICC**	**95% CI**
Whole Knee All Slices	0.25	(0-0.47)	0.22	(-0.03-0.45)	0.13	(-0.12-0.37)
Whole Knee Lateral	0.32	(0.08-0.53)	0.28	(0.03-0.49)	0.18	(-0.07-0.41)
Whole Knee Medial	0.55	(-0.07-0.41)	0.49	(0.08-0.53)	0.32	(-0.08-0.41)
Whole Knee Intercondylar	0.42	(-0.07-0.41)	0.34	(0.08-0.53)	0.17	(-0.12-0.37)
Femur All Slices	0.17	(-0.09-0.4)	0.12	(-0.14-0.36)	0.05	(-0.21-0.29)
Femur Lateral	0.31	(0.06-0.52)	0.27	(0.02-0.49)	0.16	(-0.09-0.39)
Femur Medial	0.41	(0.18-0.6)	0.37	(0.14-0.57)	0.22	(-0.03-0.44)
Femur Intercondylar	0.15	(-0.1-0.39)	0.12	(-0.14-0.36)	0.04	(-0.21-0.29)
Tibia All Slices	0.61	(0.43-0.75)	0.56	(0.36-0.71)	0.35	(0.11-0.55)
Tibia Lateral	0.39	(0.16-0.59)	0.35	(0.11-0.55)	0.22	(-0.03-0.45)
Tibia Medial	0.65	(0.48-0.78)	0.57	(0.38-0.72)	0.41	(0.18-0.6)
Tibia Intercondylar	0.65	(0.48-0.78)	0.57	(0.37-0.72)	0.33	(0.09-0.54)
Patella All Slices	0.31	(0.06-0.52)	0.32	(0.08-0.53)	0.28	(0.03-0.49)

Patella Lateral	0.50	(0.28-0.66)	0.46	(0.24-0.64)	0.39	(0.15-0.58)
Patella Medial	0.18	(-0.08-0.41)	0.12	(-0.14-0.36)	0.09	(-0.17-0.33)
Patella Intercondylar	0.37	(0.14-0.57)	0.32	(0.08-0.53)	0.27	(0.02-0.49)

Table 3.3. 2-way mixed intra-class correlation coefficients (ICC) for deep learning-generated iKIMRISS scores **in grouped neighbouring regions** compared to mean of 8 human reader scores on n=62 cases with complete human scoring data available. ICCs are shown for iKIMRISS scores calculated according to different thresholds for the percentage of BML pixels in each grid square required to assign a positive score (20%, 10%, 0%).

	20% BML pixel threshold		**10% BML pixel threshold**		**0% BML pixel threshold**	
	ICC	**95% CI**	**ICC**	**95% CI**	**ICC**	**95% CI**
FCs All Slice	0.32	(0.07-0.53)	0.29	(0.04-0.51)	0.14	(-0.12-0.39)
FCs Lateral	0.31	(0.06-0.52)	0.28	(0.02-0.5)	0.15	(-0.11-0.39)
FCs Medial	0.43	(0.19-0.62)	0.45	(0.22-0.63)	0.41	(0.17-0.6)
FCs Intercondylar	0.19	(-0.07-0.43)	0.20	(-0.05-0.44)	0.07	(-0.19-0.32)
FPs All Slice	0.32	(0.07-0.53)	0.36	(0.12-0.57)	0.12	(-0.14-0.36)

FPs Lateral	0.74	(0.59-0.84)	0.70	(0.54-0.81)	0.15	(-0.11-0.39)
FPs Medial	0.17	(-0.09-0.4)	0.13	(-0.13-0.38)	0.08	(-0.18-0.33)
FPs Intercondylar	na		na		na	
FTs All Slice	0.46	(0.23-0.64)	0.30	(0.05-0.52)	0.16	(-0.1-0.4)
FTs Lateral	0.78	(0.66-0.86)	0.69	(0.53-0.81)	0.50	(0.28-0.67)
FTs Medial	0.11	(-0.15-0.36)	0.09	(-0.17-0.33)	0.04	(-0.22-0.3)
FTs Intercondylar	0.20	(-0.06-0.43)	0.10	(-0.16-0.35)	0.08	(-0.18-0.33)
T1T11 All Slice	0.76	(0.63-0.85)	0.62	(0.43-0.75)	0.31	(0.06-0.53)
T1T11 Lateral	0.71	(0.55-0.82)	0.49	(0.27-0.67)	0.15	(-0.11-0.39)
T1T11 Medial	0.53	(0.31-0.69)	0.41	(0.17-0.6)	0.34	(0.09-0.55)
T1T11 Intercondylar	0.67	(0.5-0.79)	0.54	(0.33-0.7)	0.40	(0.16-0.6)
T2T22 All Slice	0.79	(0.67-0.87)	0.74	(0.59-0.84)	0.41	(0.17-0.6)
T2T22 Lateral	0.66	(0.48-0.78)	0.56	(0.36-0.72)	0.27	(0.02-0.49)
T2T22 Medial	0.67	(0.5-0.79)	0.61	(0.42-0.75)	0.48	(0.26-0.66)
T2T22 Intercondylar	0.70	(0.55-0.81)	0.65	(0.48-0.78)	0.36	(0.12-0.57)
T3T33 All Slice	0.67	(0.5-0.79)	0.62	(0.43-0.75)	0.50	(0.28-0.67)
T3T33 Lateral	0.41	(0.17-0.6)	0.34	(0.1-0.55)	0.29	(0.04-0.51)
T3T33 Medial	0.74	(0.6-0.84)	0.71	(0.55-0.82)	0.46	(0.23-0.64)
T3T33 Intercondylar	0.56	(0.35-0.71)	0.50	(0.28-0.67)	0.39	(0.15-0.59)

T4T44 All Slice	0.47	(0.24-0.65)	0.44	(0.21-0.63)	0.34	(0.09-0.55)
T4T44 Lateral	0.16	(-0.1-0.4)	0.14	(-0.12-0.38)	0.14	(-0.12-0.38)
T4T44 Medial	0.64	(0.45-0.77)	0.59	(0.39-0.73)	0.46	(0.23-0.64)
T4T44 Intercondylar	0.55	(0.34-0.7)	0.47	(0.24-0.65)	0.29	(0.04-0.51)
T5T55 All Slice	0.23	(-0.03-0.46)	0.22	(-0.03-0.45)	0.14	(-0.12-0.38)
T5T55 Lateral	0.11	(-0.15-0.36)	0.12	(-0.14-0.37)	0.07	(-0.19-0.32)
T5T55 Medial	0.41	(0.17-0.6)	0.28	(0.03-0.5)	0.17	(-0.09-0.41)
T5T55 Intercondylar	0.37	(0.13-0.57)	0.33	(0.08-0.54)	0.15	(-0.12-0.39)

Table 3.4. 2-way consistency intra-class correlation coefficients (ICC) for deep learning-generated iKIMRISS scores **in individual grid square regions** compared to mean of 8 human reader scores on n=62 cases with complete human scoring data available. For ease of viewing, ICCs are shown for iKIMRISS scores calculated according to 20% BML pixel threshold only.

	All Slices		Lateral		Medial		Intercondylar	
	ICC	95% CI	ICC	95% CI	ICC	95% CI	ICC	95% CI

FC1	0.19	(-0.07-0.43)	0.34	(0.09-0.55)	0.29	(0.04-0.51)	0.00	(-0.26-0.26)
FC2	0.03	(-0.23-0.29)	0.07	(-0.19-0.32)	0.08	(-0.18-0.33)	0.00	(-0.26-0.26)
FC3	0.03	(-0.23-0.28)	-0.01	(-0.27-0.25)	0.05	(-0.21-0.3)	na	
FC4	0.01	(-0.24-0.27)	0.00	(-0.25-0.26)	0.00	(-0.25-0.26)	0.00	(-0.26-0.26)
FC5	0.01	(-0.24-0.27)	-0.05	(-0.3-0.21)	0.20	(-0.06-0.43)	0.00	(-0.26-0.26)
FC6	0.08	(-0.18-0.33)	0.02	(-0.24-0.27)	0.27	(0.01-0.49)	0.00	(-0.26-0.26)
FP1	0.15	(-0.11-0.39)	0.23	(-0.03-0.46)	0.15	(-0.11-0.39)	na	
FP2	0.21	(-0.05-0.44)	0.56	(0.35-0.71)	0.13	(-0.13-0.37)	na	
FP3	0.31	(0.06-0.52)	0.74	(0.6-0.84)	0.06	(-0.2-0.31)	na	
FP4	0.13	(-0.13-0.37)	0.42	(0.18-0.61)	0.10	(-0.16-0.35)	na	
FT1	0.68	(0.52-0.8)	0.89	(0.82-0.93)	0.09	(0.17-0.34)	0.16	(-0.1-0.4)
FT2	0.06	(-0.2-0.31)	0.26	(0.01-0.49)	0.02	(-0.24-0.27)	-0.01	(-0.27-0.24)
FT3	0.27	(0.01-0.49)	0.21	(-0.05-0.44)	0.11	(-0.15-0.35)	0.37	(0.12-0.57)
FT4	0.00	(-0.26-0.26)	0.21	(-0.05-0.44)	0.11	(-0.15-0.35)	0.00	(-0.26-0.26)
T1	0.78	(0.66-0.87)	0.55	(0.35-0.71)	0.63	(0.45-0.76)	0.70	(0.54-0.81)
T2	0.81	(0.7-0.88)	0.49	(0.27-0.66)	0.69	(0.53-0.8)	0.65	(0.48-0.78)

T3	0.75	(0.61-0.85)	0.50	(0.28-0.67)	0.73	(0.59-0.83)	0.52	(0.3-0.68)
T4	0.54	(0.33-0.7)	0.32	(0.07-0.53)	0.73	(0.59-0.83)	0.48	(0.25-0.65)
T5	0.22	(-0.04-0.45)	0.19	(-0.07-0.43)	0.56	(0.35-0.71)	0.50	(0.28-0.67)
T11	0.63	(0.45-0.76)	0.77	(0.64-0.86)	0.25	(-0.01-0.47)	0.57	(0.36-0.72)
T22	0.71	(0.55-0.82)	0.69	(0.53-0.8)	0.53	(0.32-0.69)	0.74	(0.6-0.84)
T33	0.51	(0.29-0.68)	0.29	(0.04-0.51)	0.62	(0.44-0.76)	0.54	(0.32-0.7)
T44	0.35	(0.1-0.55)	0.05	(-0.21-0.3)	0.45	(0.22-0.63)	0.48	(0.25-0.65)
T55	0.16	(-0.1-0.4)	0.00	(-0.26-0.26)	0.13	(-0.13-0.37)	0.21	(-0.05-0.45)
P1	0.52	(0.3-0.68)	0.52	(0.31-0.69)	0.25	(0-0.48)	0.45	(0.21-0.63)
P2	0.69	(0.53-0.81)	0.61	(0.42-0.75)	0.18	(-0.08-0.42)	0.52	(0.31-0.69)
P3	0.47	(0.25-0.65)	0.42	(0.18-0.61)	0.07	(-0.19-0.32)	0.22	(-0.04-0.45)
P4	0.35	(0.1-0.56)	0.42	(0.18-0.61)	-0.02	(-0.27-0.24)	0.06	(-0.2-0.31)

In the large dataset, rank correlations comparing iKIMRISS scores to number of MOAKS BMLs were low to moderate at the level of the whole knee [Kendall's Tau (95% CI): 0.267 (0.226-0.308), 0.293 (0.249-0.342),and 0.333 (0.281-0.38) for all slice, lateral, and medial compartments respectively]. Much

stronger correlation occurred in the medial tibia [Tau (95% CI): 0.62 (0.58-0.664)] compared to the lateral and intercondylar tibia [[Tau (95% CI)(0.222 (0.149-0.291) and 0.319 (0.261-0.380), respectively], while femoral scores showed moderate correlation in both lateral and medial compartments [0.327 (0.281-0.377) and 0.319 (0.272-0.367), respectively]. There was essentially no correlation at all between medial patella iKIMRISS and MOAKS scores [0.06 (0.000-0.129)], but moderate correlation was seen in the lateral patella [0.356 (0.3-0.398].

Table 3.5. Rank correlations between iKIMRISS scores vs. number of Bone Marrow Lesions assigned by human MOAKS readers in n=836 patients with human reader data available.

	Kendall's Tau	**95% CI**	**p**
Whole Knee All Slice	0.267	(0.226-0.308)	<0.0001
Whole Knee Medial	0.333	(0.281-0.380)	<0.0001
Whole Knee Lateral	0.293	(0.249-0.342)	<0.0001
Femur All Slice	0.294	(0.247-0.329)	<0.0001
Femur Medial	0.319	(0.367-0.272)	<0.0001
Femur Lateral	0.326	(0.377-0.281)	<0.0001
Tibia All Slice	0.378	(0.418-0.336)	<0.0001

Tibia Medial	0.621	(0.580-0.664)	<0.0001
Tibia Lateral	0.222	(0.291-0.149)	<0.0001
Tibia Intercondyle	0.319	(0.261-0.380)	<0.0001
Patella All Slice	0.278	(0.235-0.321)	<0.0001
Patella Medial	0.0619	(0.000-0.129)	0.0074
Patella Lateral	0.356	(0.300-0.398)	<0.0001

Associations between iKIMRISS and clinical measures:

Knee replacement surgery occurred within the ten year follow-up period after the MRI occurred in 196/1631 (12.%) of cases analyzed.

Significant results from the univariable logistic regression are displayed below (Figures 3.6a-d). None of the iKIMRISS whole bone scores were found to be predictive of eventual arthroplasty, aside from a relatively small but statistically significant OR (95% CI) of 1.1 (1.0-1.1) in the medial portion of the tibia. Overall, many regions at the other two levels of scoring granularity showed weak but statistically significant ORs. These were relatively uniform across the board (ORs ranging from 1.1-1.9), with the strongest odds ratios found in the medial grid square scores for FC1 and FC3 in the weight-bearing femur [OR (95% CI): 1.7(1.4-1.9) and (1.9 (1.4-2.6), respectively] and T11 in the anterior tibia (1.7 (1.3-2.0). One significant, though weak, negative

association with eventual arthroplasty was found in the lateral patella region labeled P3 [OR (95% CI): 0.9 (0.9-1.0)].

When coronally subdivided individual grid square regions were entered into the CHAID algorithm, a combination of medial T1 (anterior tibia), lateral FP1 (posterior femur), and medial T2 (anterior tibia) scores was found to have the best predictive utility for eventual arthroplasty (Figure).

Figure 3.6a. Significant odds ratios for knee arthroplasty in n=1631 baseline MRIs from the Osteoarthritis Initiative according to deep learning-generated iKIMRISS scores in all available MRI slices. Note: Significant ORs rounding to 1.0 are displayed, but regions are not highlighted.

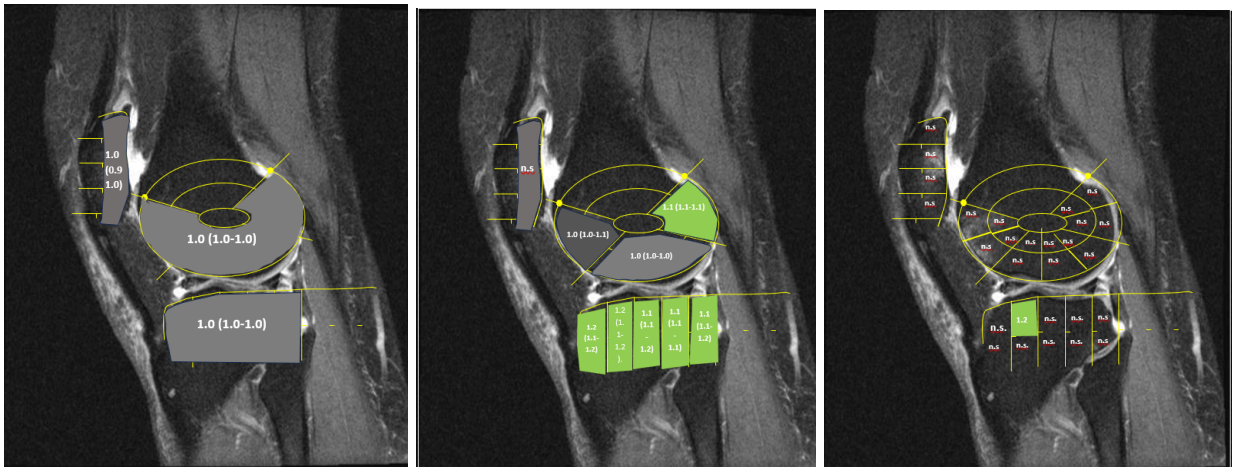


Figure 3.6b. Significant odds ratios for knee arthroplasty in n=1631 baseline MRIs from the Osteoarthritis Initiative according to deep learning-generated

iKIMRISS scores in lateral compartment of the knee joint. Note: Significant ORs rounding to 1.0 are displayed, but regions are not highlighted.

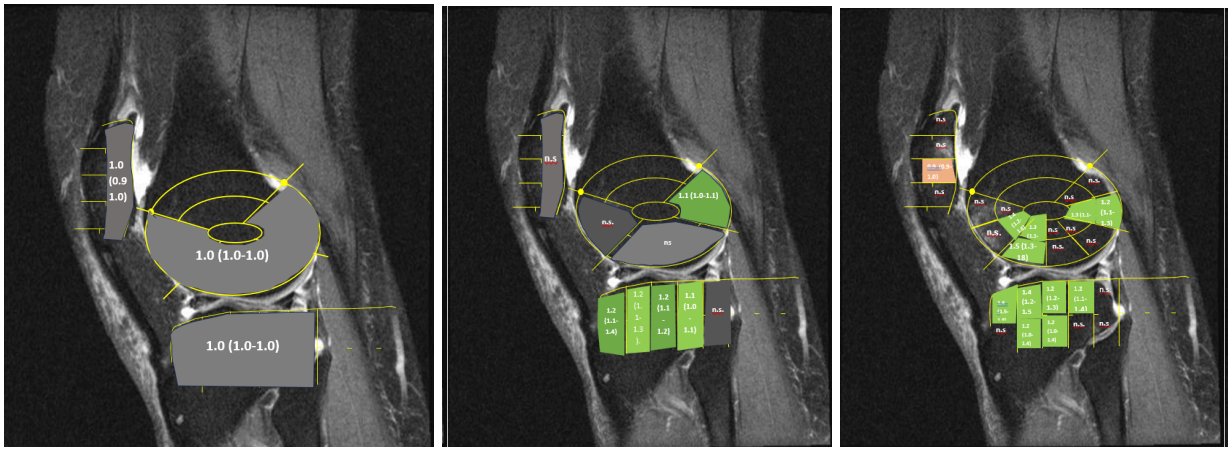


Figure 3.6c. Significant odds ratios for knee arthroplasty in n=1631 baseline MRIs from the Osteoarthritis Initiative according to iKIMRISS scores in the medial compartment of the knee joint. Note: Significant ORs rounding to 1.0 are displayed, but regions are not highlighted.

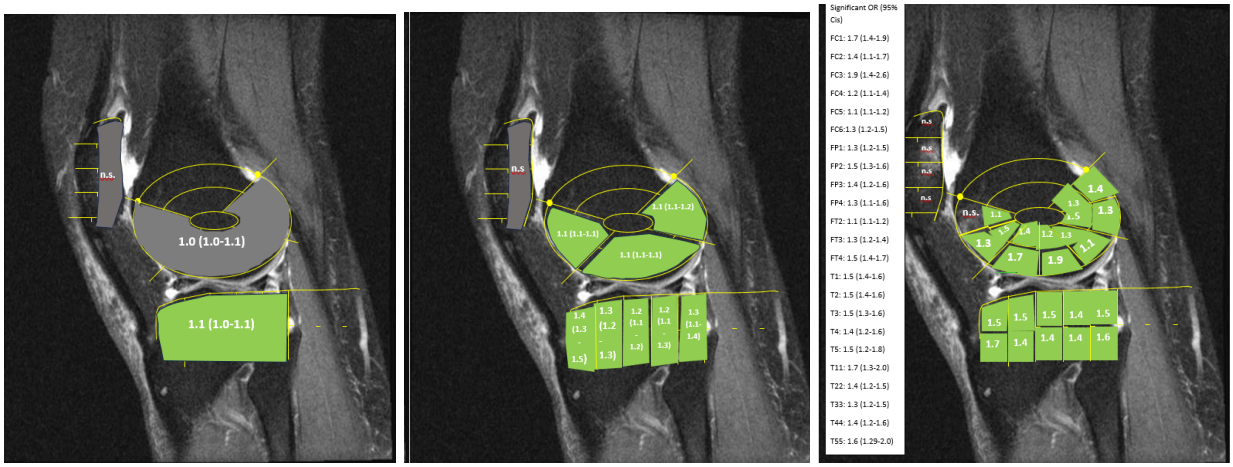


Figure 3.6d. Significant odds ratios for knee arthroplasty in n=1631 baseline MRIs from the Osteoarthritis Initiative according to deep learning-generated

iKIMRISS scores in intercondylar compartment of the knee joint. Note: Significant ORs rounding to 1.0 are displayed, but regions are not highlighted.

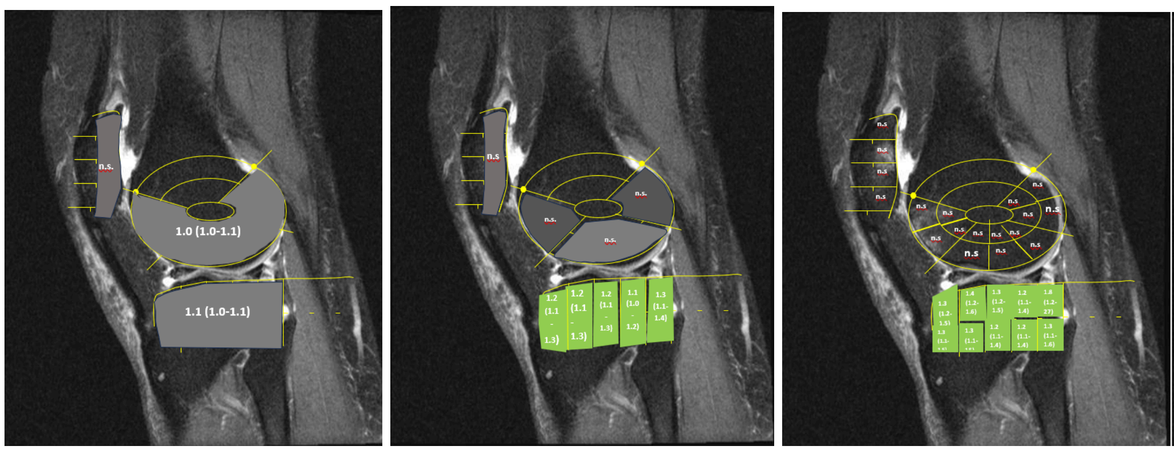
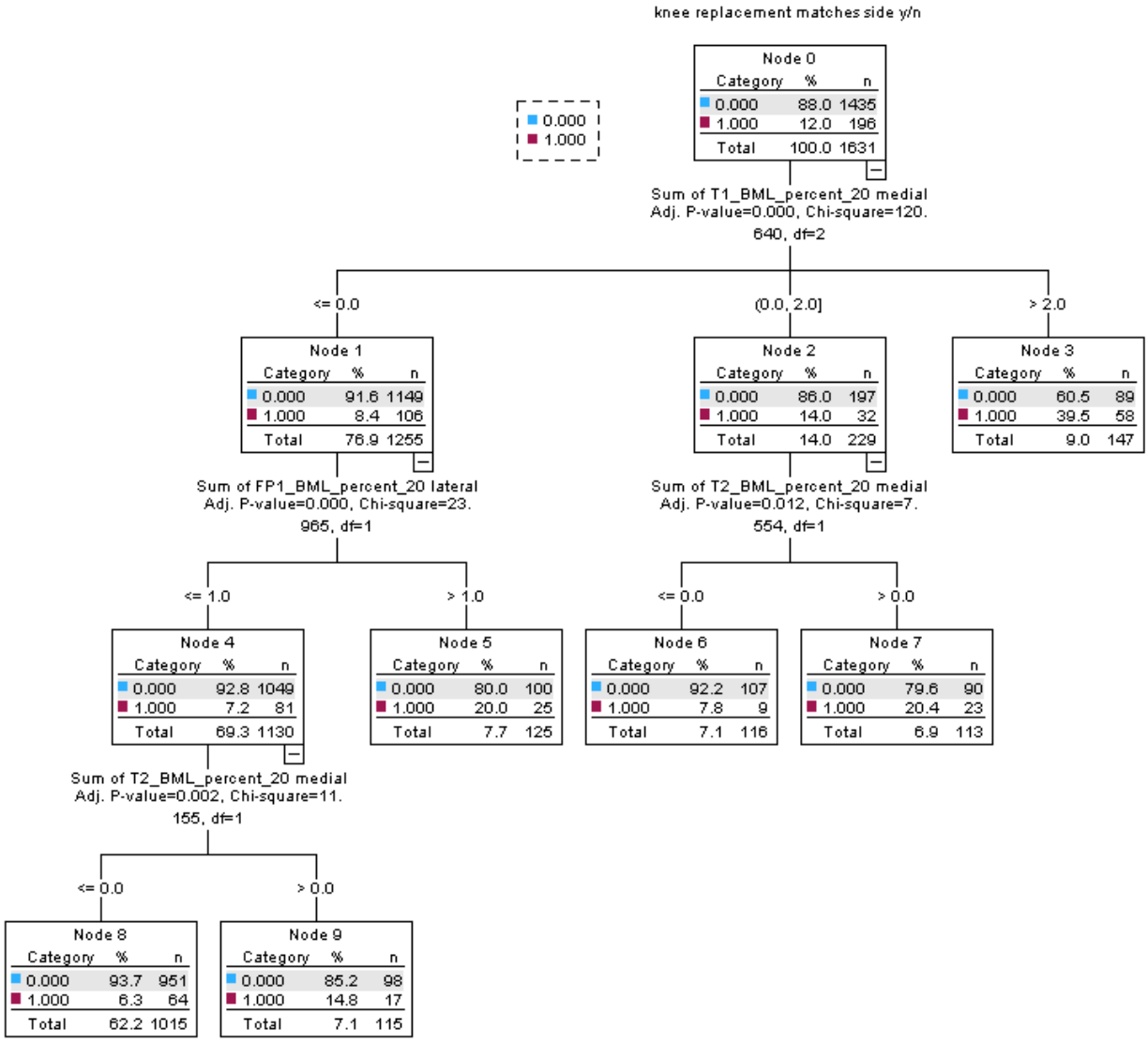


Figure 3.7. CHAID generated decision tree for eventual knee replacement considering coronally subdivided individual iKIMRISS grid square scores extracted by deep-learning.


knee replacement matches side y/n
0.000
1.000
Node 0
Category % n
0.000 88.0 1435
1.000 12.0 196
Total 100.0 1631
Sum of T1_BML_percent_20 medial
Adj. P-value=0.000, Chi-square=120.
640, df=2
<= 0.0
(0.0, 2.0]
> 2.0
Node 1
Category % n
0.000 91.6 1149
1.000 8.4 106
Total 76.9 1255
Node 2
Category % n
0.000 86.0 197
1.000 14.0 32
Total 14.0 229
Node 3
Category % n
0.000 60.5 89
1.000 39.5 58
Total 9.0 147
Sum of FP1_BML_percent_20 lateral
Adj. P-value=0.000, Chi-square=23.
965, df=1
Sum of T2_BML_percent_20 medial
Adj. P-value=0.012, Chi-square=7.
554, df=1
<= 1.0
> 1.0
<= 0.0
> 0.0
Node 4
Category % n
0.000 92.8 1049
1.000 7.2 81
Total 69.3 1130
Node 5
Category % n
0.000 80.0 100
1.000 20.0 25
Total 7.7 125
Node 6
Category % n
0.000 92.2 107
1.000 7.8 9
Total 7.1 116
Node 7
Category % n
0.000 79.6 90
1.000 20.4 23
Total 6.9 113
Sum of T2_BML_percent_20 medial
Adj. P-value=0.002, Chi-square=11.
155, df=1
<= 0.0
> 0.0
Node 8
Category % n
0.000 93.7 951
1.000 6.3 64
Total 62.2 1015
Node 9
Category % n
0.000 85.2 98
1.000 14.8 17
Total 7.1 115

Discussion

In our smaller sample where human reader KIMRISS data was available, the clinical relevance of bone marrow lesions occurring at specific small locations of the knee was apparent from analysis of scores at several different levels of granularity. While the overall incidence of BML at the level of the whole bone did not show any strong relationships to clinical measures or outcomes, significant relationships began to appear as the data for smaller subdivisions was analyzed. This was true for both relationship of BML to patient self-report measures of pain, stiffness, and disability, as well as predictive capacity for total knee replacement surgery during the follow-up period. This may indicate that there is added value in recording semi-quantitative assessment of bone marrow lesions in smaller, more granular anatomical locations rather than limiting this to broader regions.

BMLs in the medial compartment of the joint have previously been linked to to eventual arthroplasty, [62] [63] which is consistent with the results from this dataset, and our analysis demonstrated stronger ORs for this outcome when examining BML location in greater detail within the medial joint compared to larger summed regions. The same pattern was observed when analyzing WOMAC pain, stiffness, and function scores. However, due to the smaller size of the human KIMRISS dataset (n=62) and the high number of BML scoring locations analyzed, overfitting of models is quite likely, and the results from

individual coronally subdivided grid square scores may not necessarily be considered valid without further testing.

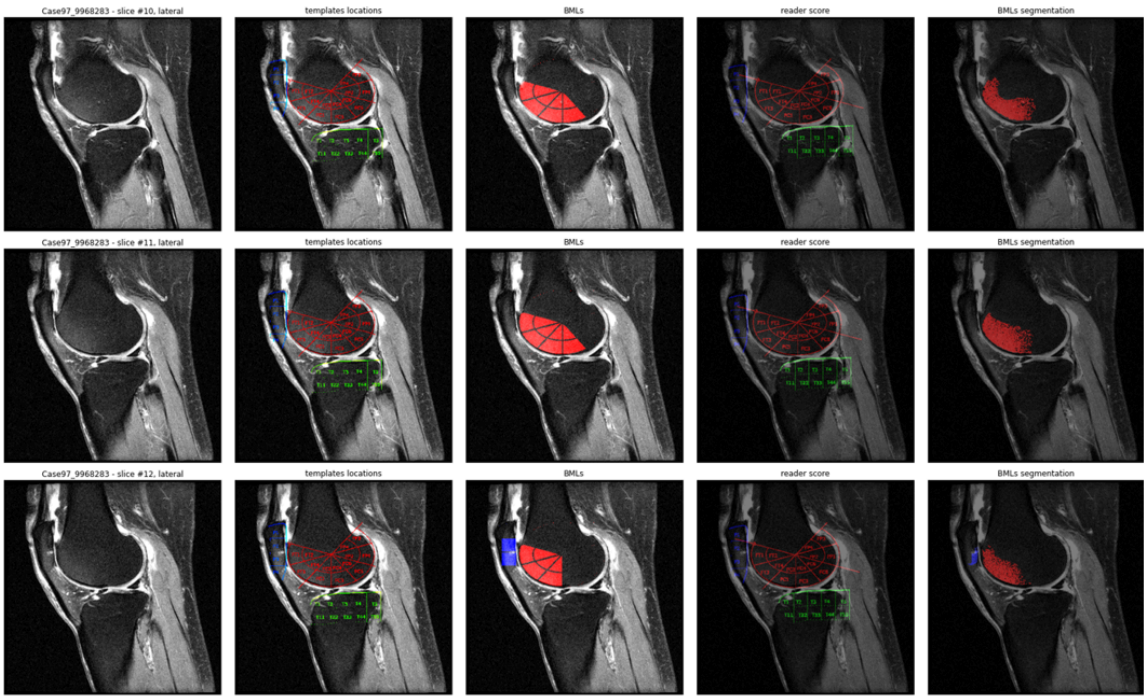
In order to repeat analyses again in a substantially larger dataset, we attempted to overcome the obstacle of expert time requirement for manual KIMRISS scoring by applying a newly developed deep-learning algorithm to n=1631 MRIs from the Osteoarthritis Initiative dataset . This algorithm, dubbed “iKIMRISS”, aims to approximate the KIMRISS scoring methodology by using automated detection of bone marrow lesions combined with automatic placement of grid location templates. The results of the clinical measure and outcome analyses of the iKIMRISS data in this large dataset showed a similar pattern of increased significance of BML presence in smaller specific regions as scoring granularity increased from the level of the whole bone to the level of coronally-subdivided grid regions, though effect sizes were generally weaker and there were fewer stand-out regions of interest.

Despite a robust sample size that should should theoretically produce results translatable to the broader osteoarthritis population, the quality of the BML data extracted by the current version of the algorithm presents extensive limitations. Its reliability when compared to the small dataset of human expert KIMRISS scores was spotty and significantly poorer across all levels of granularity compared to reliability previously demonstrated between different human readers. In the absence of human KIMRISS scores for comparison, correlations to the somewhat analogous MOAKS BML scores across coronal

subdivisions of the femur, tibia, and patella were not particularly strong, and showed similar reliability patterns to those demonstrated by the ICC analysis in the small dataset. Scores in the femur were particularly inconsistent with human reader assessment by either methodology, which is notable considering this tends to be an area of higher reliability amongst human reader pairs.

This lack of reliability suggests that the iKIMRISS algorithm requires additional training in order to more closely approximate human scores, even in the most broadly defined regions. While it is time-consuming and beyond the scope of this study to examine every single MRI slice analyzed by the algorithm, spot-checks of the algorithm's output superimposed onto the images show promising segmentation of BML. However, there are also instances of failure to distinguish hyperintense signal representing a true lesion vs. hyperintense signal caused by various types of artifact [64]. An example of this can be seen in the figure below, where an area of slightly increased signal in the femur has resulted in the assignment of positive BML scores in many grid squares in the femur across several slices, despite unequivocal agreement by expert human readers that no lesion is present in the femur (Figure).

Figure 3.8. An example of widespread false positive lesion assignment by deep-learning in the femur (rows 1,2, and 3) and patella (row 3).



We attempted to minimize discrepancies using the data output from the current iteration of the algorithm, including imposing a threshold for the percentage of a grid square occupied by BML pixels required to assign a positive score. This would prevent overscoring caused by the automated assignment of a positive score from a very small cluster of BML pixels that may be overlooked by a human reader. The algorithm's probability threshold for BML classification was also adjusted after reviewing instances of false negative and false positive scores in order to optimize reliability. However, the results of our reliability analyses show that these measures alone were not enough to achieve adequate reliability with human reader scores. At present, we can not be sure that signal is outweighing noise when it comes to these deep

learning-generated scores, even at the lowest level of granularity. The algorithm is still in the early stages of development, and while it is unfortunate that we can not yet draw conclusions from these analyses, the preliminary results do make a case for continuing to refine the algorithm. The next step is to provide more ground truth data (human generated BML segmentation and grid template placements) from a wide variety of cases to the algorithm for additional training so it may navigate different presentations of the knee–including those containing artifact–while producing BML scores more comparable to those generated by humans. Once this is complete to the point of achieving high reliability, more complete and readily-interpretable analyses can be carried out on this large clinical dataset.

CONCLUSIONS AND FURTHER WORK

Previous work has shown that semi-quantitative scoring systems capturing more granular detail can improve inter-rater reliability when all granular observations are added together into a total score. This is supported in the opening chapter by the increased similarity seen between decision-trees from multiple readers' granular axSpA SIJ scoring data when compared to global scoring data. It could be that more detailed, rigid scoring frameworks eliminate some subjectivity and more sharply focus readers' assessments, allowing for systematic capture of small differences or changes overall.

However, when analyzed individually, isolated granular observations may present with inter-rater reliability challenges. Certain discrepancies between readers, such as slice selection or grid template placement, may be somewhat negated when all scores are lumped together, becoming apparent only when finer distinctions are analyzed (eg. separating medial slices from intercondylar slices, weight-bearing region of the femur from posterior femur, etc.). As shown in the KIMRISS human reader reliability chapter, certain anatomical locations may be more affected by inter-rater disagreement than others. If there is a possibility that individual granular components of a score (e.g. BML presence in certain KIMRISS grid square locations on certain slices) serve as significant clinical predictors or correlates, it is important to search for sources of discrepancy within these components and attempt to calibrate readers or algorithms for increased reliability on a granular level.

Moving forward, the sort of granular reliability analyses carried out in this thesis could be used to inform adjustments to scoring methodology or training methods, both for human readers and automated systems. While this work has dealt only with single timepoint status scores, it would be particularly beneficial to perform the same analyses on longitudinal change data, as change capture is one very important function of semi-quantitative methodology and may present with different reliability challenges and potential for utility.

References

1. Abhishek, A. & Doherty, M. Diagnosis and clinical presentation of osteoarthritis. *Rheum. Dis. Clin. North Am.* **39**, 45–66 (2013).
2. Zhang, W. *et al.* EULAR evidence-based recommendations for the diagnosis of knee osteoarthritis. *Ann. Rheum. Dis.* **69**, 483–489 (2010).
3. Felson, D. T. & Hodgson, R. Identifying and treating preclinical and early osteoarthritis. *Rheum. Dis. Clin. North Am.* **40**, 699–710 (2014).
4. Rudwaleit, M., Khan, M. A. & Sieper, J. The challenge of diagnosis and classification in early ankylosing spondylitis: do we need new criteria? *Arthritis Rheum.* **52**, 1000–1008 (2005).
5. Lassiter, W., Bhutta, B. S. & Allam, A. E. *Inflammatory Back Pain*. (StatPearls Publishing, 2024).

6. Sen, R. & Hurley, J. A. *Osteoarthritis*. (StatPearls Publishing, 2023).
7. Dobson, F. *et al.* OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. *Osteoarthritis Cartilage* **21**, 1042–1052 (2013).
8. Vesović-Potić, V., Mustur, D., Stanisavljević, D., Ille, T. & Ille, M. Relationship between spinal mobility measures and quality of life in patients with ankylosing spondylitis. *Rheumatol. Int.* **29**, 879–884 (2009).
9. Baraliakos, X., Szumski, A., Koenig, A. S. & Jones, H. The role of C-reactive protein as a predictor of treatment response in patients with ankylosing spondylitis. *Semin. Arthritis Rheum.* **48**, 997–1004 (2019).
10. Queiroga, F. *et al.* A scoping review of patient self-report measures of flare in knee and hip osteoarthritis (OA): A report from the OMERACT flares in OA working group. *Semin. Arthritis Rheum.* **63**, 152281 (2023).
11. van der Heijde, D. & Landewé, R. Chapter 15 - Assessment of Disease Activity, Function, and Quality of Life. in *Ankylosing Spondylitis and the Spondyloarthropathies* (eds. Weisman, M. H., van der Heijde, D. & Reveille, J. D.) 206–213 (Mosby, Philadelphia, 2006).
12. Altman, R. *et al.* Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. Diagnostic and Therapeutic Criteria Committee of the American Rheumatism Association. *Arthritis Rheum.* **29**, 1039–1049 (1986).
13. van der Linden, S., Valkenburg, H. A. & Cats, A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New

York criteria. *Arthritis Rheum.* **27**, 361–368 (1984).

14. van der Heijde, D. *et al.* Modified stoke ankylosing spondylitis spinal score as an outcome measure to assess the impact of treatment on structural progression in ankylosing spondylitis. *Rheumatology* **58**, 388–400 (2019).
15. Ostergaard, M. & Lambert, R. G. W. Imaging in ankylosing spondylitis. *Ther. Adv. Musculoskelet. Dis.* **4**, 301–311 (2012).
16. Issın, A., Öner, A., Koçkara, N. & Özcan, S. Shortest time interval for detecting the progression of knee osteoarthritis on consecutive X-rays. *Turk J Phys Med Rehabil* **66**, 383–387 (2020).
17. Al-Mnayyis, A. 'a *et al.* Radiological Insights into Sacroiliitis: A Narrative Review. *Clin. Pract.* **14**, 106–121 (2024).
18. Guermazi, A. *et al.* Prevalence of abnormalities in knees detected by MRI in adults without knee osteoarthritis: population based observational study (Framingham Osteoarthritis Study). *BMJ* **345**, e5339 (2012).
19. Hobby, J. L., Tom, B. D., Todd, C., Bearcroft, P. W. & Dixon, A. K. Communication of doubt and certainty in radiological reports. *Br. J. Radiol.* **73**, 999–1001 (2000).
20. Khorasani, R. *et al.* Is terminology used effectively to convey diagnostic certainty in radiology reports? *Acad. Radiol.* **10**, 685–688 (2003).
21. D'Agostino, M. A. *et al.* Improving domain definition and outcome instrument selection: Lessons learned for OMERACT from imaging. *Semin. Arthritis Rheum.* **51**, 1125–1133 (2021).
22. Lynch, J. A. *et al.* Comparison of BLOKS and WORMS scoring systems part

I. Cross sectional comparison of methods to assess cartilage morphology, meniscal damage and bone marrow lesions on knee MRI: data from the osteoarthritis initiative. *Osteoarthritis Cartilage* **18**, 1393–1401 (2010).

23. Hunter, D. J. *et al.* Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score). *Osteoarthritis Cartilage* **19**, 990–1002 (2011).
24. Jaremko, J. L. *et al.* Preliminary validation of the Knee Inflammation MRI Scoring System (KIMRISS) for grading bone marrow lesions in osteoarthritis of the knee: data from the Osteoarthritis Initiative. *RMD Open* **3**, e000355 (2017).
25. Maksymowych, W. P. *et al.* Comparative validation of the knee inflammation MRI scoring system and the MRI osteoarthritis knee score for semi-quantitative assessment of bone marrow lesions and synovitis-effusion in osteoarthritis: an international multi-reader exercise. *Ther. Adv. Musculoskelet. Dis.* **15**, 1759720X231171766 (2023).
26. Lin, Y. *et al.* Deep Learning Algorithm of the SPARCC Scoring System in SI Joint MRI. *J. Magn. Reson. Imaging* (2024) doi:10.1002/jmri.29211.
27. Ożga, J. *et al.* Performance of Fully Automated Algorithm Detecting Bone Marrow Edema in Sacroiliac Joints. *J. Clin. Med. Res.* **12**, (2023).
28. Felfeliyan, B., Hareendranathan, A., Kuntze, G., Jaremko, J. & Ronsky, J. Toward accurate MRI bone and cartilage segmentation in small data sets via an improved mask RCNN: data from the osteoarthritis initiative. *Osteoarthritis Cartilage* **29**, S349–S350 (2021).

29. Felfeliyan, B. *et al.* OMERACT validation of a deep learning algorithm for automated absolute quantification of knee joint effusion versus manual semi-quantitative assessment. *Semin. Arthritis Rheum.* **66**, 152420 (2024).
30. Beaton D, Maxwell L ,Grosskleg S, Shea B, Tugwell B (editors). *The OMERACT Handbook. Chapter 5: Instrument Selection for Core Outcome Sets*. https://omeract.org/handbook/ (2021).
31. Kass, G. V. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *J. R. Stat. Soc. Ser. C Appl. Stat.* **29**, 119–127 (1980).
32. Schelhorn, J. *et al.* Application of classification trees for the qualitative differentiation of focal liver lesions suspicious for metastasis in gadolinium-EOB-DTPA-enhanced liver MR imaging. *Rofo* **184**, 788–794 (2012).
33. Lang, N., Su, M.-Y., Xing, X., Yu, H. J. & Yuan, H. Morphological and dynamic contrast enhanced MR imaging features for the differentiation of chordoma and giant cell tumors in the Axial Skeleton. *J. Magn. Reson. Imaging* **45**, 1068–1075 (2017).
34. Devlin, H. *et al.* Diagnosing osteoporosis by using dental panoramic radiographs: the OSTEODENT project. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* **104**, 821–828 (2007).
35. Maksymowych, W. P. *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.* **78**, 1550–1558 (2019).
36. Lorenzin, M. *et al.* Spine and Sacroiliac Joints Lesions on Magnetic

Resonance Imaging in Early Axial-Spondyloarthritis During 24-Months Follow-Up (Italian Arm of SPACE Study). *Front. Immunol.* **11**, 936 (2020).

37. de Hooge, M. *et al.* Patients with chronic back pain of short duration from the SPACE cohort: which MRI structural lesions in the sacroiliac joints and inflammatory and structural lesions in the spine are most specific for axial spondyloarthritis? *Ann. Rheum. Dis.* **75**, 1308–1314 (2016).
38. de Winter, J. *et al.* Magnetic Resonance Imaging of the Sacroiliac Joints Indicating Sacroiliitis According to the Assessment of SpondyloArthritis international Society Definition in Healthy Individuals, Runners, and Women With Postpartum Back Pain. *Arthritis Rheumatol* **70**, 1042–1048 (2018).
39. Ortolan, A. *et al.* Are gender-specific approaches needed in diagnosing early axial spondyloarthritis? Data from the SPondyloArthritis Caught Early cohort. *Arthritis Res. Ther.* **20**, 218 (2018).
40. Lorenzin, M. *et al.* Relationship between sex and clinical and imaging features of early axial spondyloarthritis: results from a 48 month follow-up (Italian arm of the SPondyloArthritis Caught Early (SPACE) study). *Scand. J. Rheumatol.* **52**, 519–529 (2023).
41. Roemer, F. W. *et al.* MRI-detected subchondral bone marrow signal alterations of the knee joint: terminology, imaging appearance, relevance and radiological differential diagnosis. *Osteoarthritis Cartilage* **17**, 1115–1131 (2009).
42. Shi, X. *et al.* Bone marrow lesions in osteoarthritis: From basic science to

clinical implications. *Bone Rep* **18**, 101667 (2023).

43. Felson, D. T. *et al.* The association of bone marrow lesions with pain in knee osteoarthritis. *Ann. Intern. Med.* **134**, 541–549 (2001).
44. Zhang, Y. *et al.* Fluctuation of knee pain and changes in bone marrow lesions, effusions, and synovitis on magnetic resonance imaging. *Arthritis Rheum.* **63**, 691–699 (2011).
45. Fan, T. *et al.* The interactions between MRI-detected osteophytes and bone marrow lesions or effusion-synovitis on knee symptom progression: an exploratory study. *Osteoarthritis Cartilage* **29**, 1296–1305 (2021).
46. Kijowski, R., Stanton, P., Fine, J. & De Smet, A. Subchondral bone marrow edema in patients with degeneration of the articular cartilage of the knee joint. *Radiology* **238**, 943–949 (2006).
47. Kazakia, G. J. *et al.* Bone and cartilage demonstrate changes localized to bone marrow edema-like lesions within osteoarthritic knees. *Osteoarthritis Cartilage* **21**, 94–101 (2013).
48. Hafezi-Nejad, N., Zikria, B., Eng, J., Carrino, J. A. & Demehri, S. Predictive value of semi-quantitative MRI-based scoring systems for future knee replacement: data from the osteoarthritis initiative. *Skeletal Radiol.* **44**, 1655–1662 (2015).
49. Peterfy, C. G. *et al.* Whole-Organ Magnetic Resonance Imaging Score (WORMS) of the knee in osteoarthritis. *Osteoarthritis Cartilage* **12**, 177–190 (2004).
50. Hunter, D. J. *et al.* 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.* **67**, 206–211 (2008).

51. Maksymowych, W. P. *et al.* The OMERACT Knee Inflammation MRI Scoring System: Validation of quantitative methodologies and tri-compartmental overlays in osteoarthritis. *Semin. Arthritis Rheum.* **51**, 925–928 (2021).
52. deSouza, N. M. *et al.* Validated imaging biomarkers as decision-making tools in clinical trials and routine practice: current status and recommendations from the EIBALL* subcommittee of the European Society of Radiology (ESR). *Insights Imaging* **10**, 87 (2019).
53. Maksymowych, W. P. MRI and X-ray in axial spondyloarthritis: the relationship between inflammatory and structural changes. *Arthritis Res. Ther.* **14**, 207 (2012).
54. Raynauld, J.-P. *et al.* Risk factors predictive of joint replacement in a 2-year multicentre clinical trial in knee osteoarthritis using MRI: results from over 6 years of observation. *Ann. Rheum. Dis.* **70**, 1382–1388 (2011).
55. Giraudeau, B., Mallet, A. & Chastang, C. Case Influence on the Intraclass Correlation Coefficient Estimate. *Biometrics* **52**, 1492–1497 (1996).
56. Pang, J. *et al.* Quantification of bone marrow lesion volume and volume change using semi-automated segmentation: data from the osteoarthritis initiative. *BMC Musculoskelet. Disord.* **14**, 3 (2013).
57. Walsh, D. A., Sofat, N., Guermazi, A. & Hunter, D. J. Osteoarthritis Bone Marrow Lesions. *Osteoarthritis Cartilage* **31**, 11–17 (2023).

58. Scher, C., Craig, J. & Nelson, F. Bone marrow edema in the knee in osteoarthrosis and association with total knee arthroplasty within a three-year follow-up. *Skeletal Radiol.* **37**, 609–617 (2008).

59. Yusuf, E., Kortekaas, M. C., Watt, I., Huizinga, T. W. J. & Kloppenburg, M. Do knee abnormalities visualised on MRI explain knee pain in knee osteoarthritis? A systematic review. *Ann. Rheum. Dis.* **70**, 60–67 (2011).

60. Bellamy, N., Buchanan, W. W., Goldsmith, C. H., Campbell, J. & Stitt, L. W. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J. Rheumatol.* **15**, 1833–1840 (1988).

61. Liljequist, D., Elfving, B. & Skavberg Roaldsen, K. Intraclass correlation - A discussion and demonstration of basic features. *PLoS One* **14**, e0219854 (2019).

62. Tanamas, S. K. *et al.* Bone marrow lesions in people with knee osteoarthritis predict progression of disease and joint replacement: a longitudinal study. *Rheumatology* **49**, 2413–2419 (2010).

63. Roemer, F. W. *et al.* Large bone marrow lesions and worsening of bone marrow lesions in the medial tibio-femoral compartment are associated with knees undergoing total knee replacement : data from the osteoarthritis initiative. *Osteoarthritis Cartilage* **20**, S33–S34 (2012).

64. Schilling, J. H., Miro, P. & Chan, B. Y. Normal variants, imaging artifacts, and other diagnostic pitfalls in articular cartilage imaging of the

extremities. *Journal of Cartilage & Joint Preservation* **4**, 100147 (2024).