Intra- and inter-observer reproducibility of meniscal damage scores at 3.0T: using WORMS

S. Balamoody[1], M. Oldale[1], and C. E. Hutchinson[1]
[1]University of Manchester, Manchester, United Kingdom

Background: Several semi-quantitative scoring systems for osteoarthritis (OA) assessment have been developed which take into account multiple pathological features in important components of the knee joint. Advantages include providing a more comprehensive assessment of the knee joint as well as requiring fewer resources compared with fully quantitative methods. One of the most cited is the WORMS system [1]. Validation for inter-observer and intra-observer variability is essential and important in the context of multi-centre trials where multiple observers are likely. Studies which have used these systems have been validated by scoring images acquired at 1.5 T. The observers employed are often highly experienced in musculoskeletal radiology and with the scoring system itself. The aim of this study was to investigate the reproducibility of scoring systems for OA for images acquired at 3.0 Tesla. A sub-aim was to investigate the advantages and disadvantages of three different MR sequences in detection of meniscal tears.

Method: 50 consecutive knee scans were randomly selected from the database of clinical knee scans performed in 2005 on the 3.0T MR scanner at Hope Hospital, Manchester, UK. Of these, 13 subjects were found to have had the three sequences chosen to be investigated. Subjects were aged 41.0±9.5 years (range 27-64 yrs) and had a range of diagnoses including acute and chronic pathology and post-surgical follow-up. All images were acquired in a 3.0 Tesla superconducting magnet (Philips Medical Systems) using a dedicated 8-channel phased array coil. Acquisition parameters for the three sequences (sagittal plane only) were as follows:
Proton Density-weighted SPAIR - TR 1786ms, TE 20ms, slice thickness 2.5mm, slice gap 0.8mm, echo train length 7, FOV 15cm x 15cm, acquisition matrix 320 x 256; **T2-weighted TSE** - TR 3469.7, TE 100, slice thickness 2.5mm slice gap 0.8, echo train length 19, FOV 15cm x 15cm, acquisition matrix 304 x 241; **3DWATSc** - TR 20ms, TE 7.5ms, Slice thickness 3mm, Slice gap -1.5mm, echo train length 0, FOV 15cm x15cm, acquisition matrix 304 x 304.
Scans were scored independently by 3 observers, giving a separate score for each of the 3 sequences using the WORMS system. Observers had a range of experience and comprised a consultant musculoskeletal radiologist (observer B), a senior radiology registrar (observer C) and a junior radiology registrar undertaking research in OA (observer A). Observer A scored the scans twice (A(1) & A(2)) in order to assess intra-observer variability. Each sub-region of the meniscus was graded (anterior horn, posterior horn and body) and a weighted cumulative score given for each medial and lateral meniscus as described in the original WORMS publication.

Table 1: Frequency of cumulative scores given by the observers for each MR sequence

Cumul-ative Score	Frequency of scores by observers											
	PDW SPAIR				T2W				3DW			
	A (1)	A(2)	B	C	A (1)	A(2)	B	C	A (1)	A(2)	B	C
0	13	11	14	18	15	13	15	20	17	15	16	21
1	5	6	3	2	4	4	3	1	3	2	2	2
2	1	3	3	2	0	3	2	1	0	4	3	1
3	1	1	1	0	1	1	1	1	2	1	0	0
4	2	2	1	1	2	2	1	0	2	1	1	0
5	3	0	2	1	3	0	2	1	1	1	2	0
6	1	3	2	2	1	3	2	2	1	2	2	2

Results: Table 1 shows the frequency of cumulative meniscal grades for the 13 subjects by sequence and by observer. All of the observers detected the most abnormalities on the PDW SPAIR sequence and the least abnormalities on the 3DWATSc. Observer C detected fewer abnormalities compared to observers A and B. Kappa statistics were performed for combinations of observers to assess reproducibility. For this, the scores were first dichotomised to form the following categories: 0 = no tear; 1 = tear present (ie scores 1-6).

Table 2 displays results for each sequence. Looking at the results for agreement of observers between sequences, the T2W sequence consistently gave better inter- and intra- observer agreement with 50% values > 0.6 and 33% close at 0.58. The 3DWATSc sequence gave the second highest results for 4 out of 6 observer combinations. This sequence was also the least sensitive for all observers which may partially contribute to this result. The best agreement was between observers A(2) and B (consultant and junior research fellow-second score).

Table 2: Kappa values for inter/intra observer agreement

Observers	Kappa value		
	PD SPAIR	T2W	3D W ATSc
A(1) vs A(2)	0.69	0.69	0.68
A(1) vs B	0.46	0.68	0.59
A(1) vs C	0.46	0.58	0.43
A(2) vs B	0.62	0.84	0.76
A(2) vs C	0.35	0.46	0.49
B vs C	0.52	0.58	0.55

Discussion: Striking a balance between precision, reproducibility and producing a meaningful score in terms of clinical severity is challenging. The results show that in assessment of menisci, inter- and intra observer reproducibility is sequence-dependent, with T2-weighted sequences preferable. Good inter-observer agreement requires focussed training of the observers in assessing menisci and grading. Use of a more simplified scoring system may improve reproducibility as variability is noted in scoring grade 1 radial/parrot beak tears in particular.

References: 1. Peterfy C.G. et al, *Whole-Organ Magnetic Resonance Imaging Score (WORMS) of the knee in osteoarthritis*, Osteoarthritis Cartilage, 2004, 12(3), 177-90. **Acknowledgements:** Translational Imaging Unit (University of Manchester) for funding imaging.