European Radiology (2019) 29:4990–4998
https://doi.org/10.1007/s00330-019-06047-5

MUSCULOSKELETAL

Facet arthropathy evaluation: CT or MRI?

Linda Berg[1,2] · Hanne Thoresen[1] · Gesche Neckelmann[3] · Håvard Furunes[4,5,6] · Christian Hellum[7] · Ansgar Espeland[3,8]

Received: 29 August 2018 / Revised: 31 December 2018 / Accepted: 25 January 2019 / Published online: 22 February 2019

Abstract

Objective To assess the reliability of lumbar facet arthropathy evaluation with computed tomography (CT) or magnetic resonance imaging (MRI) in patients with and without lumbar disc prosthesis and to estimate the reliability for individual CT and MRI findings indicating facet arthropathy.

Methods Metal-artifact reducing CT and MRI protocols were performed at follow-up of 114 chronic back pain patients treated with ($n = 66$) or without ($n = 48$) lumbar disc prosthesis. Three experienced radiologists independently rated facet joint space narrowing, osteophyte/hypertrophy, erosions, subchondral cysts, and total grade facet arthropathy at each of the three lower lumbar levels on both CT and MRI, using Weishaupt et al's rating system. CT and MRI examinations were randomly mixed and rated independently. Findings were dichotomized before analysis. Overall kappa and (due to low prevalence) prevalence- and bias-adjusted kappa were calculated to assess interobserver agreement.

Results Interobserver agreement on total grade facet arthropathy was moderate at all levels with CT (kappa 0.47–0.48) and poor to fair with MRI (kappa 0.20–0.32). Mean prevalence- and bias-adjusted kappa was lower for osteophyte/hypertrophy versus other individual findings (CT 0.58 versus 0.79–0.86, MRI 0.35 versus 0.81–0.90), higher with CT versus MRI when rating osteophyte/hypertrophy (0.58 versus 0.35) and total grade facet arthropathy (0.54 versus 0.31), and generally similar at levels with versus levels without disc prosthesis.

Conclusion Interobserver agreement on facet arthropathy was moderate with CT and better with CT than with MRI. Disc prosthesis did not influence agreement. A more reliable grading of facet arthropathy requires a more consistent evaluation of osteophytes/hypertrophy.

Key Points

- *In this study, interobserver agreement on facet arthropathy (FA) severity—based on facet joint space narrowing, osteophyte/hypertrophy, erosions, and subchondral cysts—was better with CT versus MRI.*
- *Metal-artifact reducing CT and MRI protocols helped to improve visibility and maintain agreement when evaluating severity of FA at levels with metallic disc prosthesis.*
- *Agreement was poorer for severity of osteophytes/hypertrophy than for the other evaluated FA findings; improved agreement on total grade FA evaluated with CT or MRI thus requires more consistent grading of osteophytes/hypertrophy between different radiologists.*

Keywords Reproducibility of results · Magnetic resonance imaging · Tomography, X-ray computed · Zygapophyseal joint · Prostheses and implants

✉ Linda Berg
linda.berg@nlsh.no

[1] Department of Radiology, Nordland Hospital, Postbox 1480, 8092 Bodø, Norway

[2] Department of Clinical Medicine, Faculty of Health Sciences, University of Tromsø, 9037 Tromsø, Norway

[3] Department of Radiology, Haukeland University Hospital, Jonas Liesvei 65, 5021 Bergen, Norway

[4] Department of Surgery, Innlandet Hospital Gjøvik, Kyrre Grepps gate 11, 2819 Gjøvik, Norway

[5] University of Oslo, Postbox 1072 Blindern, 0316 Oslo, Norway

[6] Oslo University Hospital Ullevål, FORMI, Building 37B, Postbox 4950 Nydalen, 0424 Oslo, Norway

[7] Division of Orthopaedic Surgery, Oslo University Hospital, Postbox 4950 Nydalen, 0424 Oslo, Norway

[8] Department of Clinical Medicine, University of Bergen, Pb 7804, 5020 Bergen, Norway

Abbreviations

BW	Pixel bandwidth
CT	Computed tomography
DICOM	Digital Imaging and Communications in Medicine
ETL	Echo train length
FA	Facet arthropathy
MRI	Magnetic resonance imaging
NEX	Number of excitations
PABAK	Prevalence- and bias-adjusted kappa
PACS	Picture Archiving and Communication System
TE	Echo time
TR	Repetition time

Introduction

Facet arthropathy (FA) is prevalent in patients with low back pain [1, 2] and has been studied for an association with pain and for a potential impact on treatment indications and outcome [3]. Many studies have not revealed any association between pain and presence or severity of FA on computed tomography (CT) or magnetic resonance imaging (MRI) [2, 4]. However, results for prevalence and impact of FA are diverging. In a systematic review, the threshold for diagnosis of FA varied between studies and estimates of prevalence in people without low back pain ranged from 3 to 76% [5]. Furthermore, FA can affect spinal motion and stability [2, 6]. This is regarded important in degenerative spondylolisthesis and in candidates for disc prosthesis surgery [7–10]. FA may increase at the prosthesis level after the surgery [10, 11], and the increase was associated with worse outcome in one study [12].

Before we can conclude regarding associations between imaging findings and clinical findings, both types of findings must first demonstrate adequate reliability [13, 14]. It has been stated that most studies did not use reliable scales for rating of FA [2] and that better rating criteria and more consistent interpretations are needed [15]. Interobserver reliability for FA was often poor to moderate [15–18]. Unreliable assessment of FA may lead to diverging and faulty conclusions regarding the relevance of FA for pain, treatment choice, prognosis, and treatment effect.

To improve the rating of FA, we need to know the reliability for each finding which the overall FA grade is based on, but no study has examined this. In particular, we should optimize the rating of FA in patients with lumbar disc prosthesis. FA may increase at the prosthesis level [10, 11], but can be difficult to evaluate at the prosthesis level, as the prosthesis causes metal artifacts on CT and MRI. This study applied metal-artifact reducing CT and MRI protocols. The aim of the study was to assess the reliability of lumbar FA evaluation with CT or MRI in patients with and without lumbar disc prosthesis and to estimate the reliability for individual MRI and CT findings indicating FA.

Materials and methods

The present study was approved by the Norwegian Regional Ethical Committee South East C (2011/2177). All patients gave their informed consent prior to inclusion.

Patients

This retrospective reliability study was based on 8-year follow-up imaging of 114 of 173 patients randomized to disc prosthesis surgery or non-operative treatment in a prospective multicenter trial in 2004–2007 [19]. Patients eligible for the trial were 25–55 years, had chronic LBP, and had disc degeneration at L4/L5 and/or L5/S1 on MRI. Patients were excluded if they had disc degeneration at any higher level (L1–L4) or had spondylolysis, spondylolisthesis, arthritis, osteoporosis, prior fracture L1-S1, prior spinal fusion, deformity, or symptomatic disc herniation/spinal stenosis [19]. FA was not an exclusion criterion. Patients eligible for this reliability study had completed both CT and MRI at 8-year follow-up according to standardized protocols (see below). We excluded 10 patients who had undergone spinal fusion during follow-up and 14 (of 128) patients with both CT and MRI because non-study imaging protocols had been used.

Imaging

CT and MRI were conducted on the same day from August 2012 to August 2015, using metal-artifact reducing protocols with continuous axial images parallel to the L4/L5 disc, covering the three lower lumbar levels. Multi-slice CT was performed on different CT scanners based on the following protocol: 140 kVp, reference mAs 200, soft reconstruction kernel (B20), and 3-mm-thick axial, sagittal, and coronal reformatted slices [20]. MRI was performed on the same type of 1.5-T units at different institutions using thin slices (3 mm) and increased pixel bandwidth (BW), echo train length (ETL), and number of excitations (NEX) [20, 21]. All MRI examinations included (a) sagittal T1-weighted fast spin echo images with repetition time (TR)/echo time (TE) 549–610 ms/7.7–8 ms, matrix 448 × 448, BW 698, ETL 7, NEX 4; (b) sagittal T2-weighted fast spin echo images with TR 4480–5000 ms/TE 92–94 ms, matrix 448 × 448, BW 413, ETL 30, NEX 3; and (c) axial T2 fast spin echo images with TR 4201–5190 ms/TE 72–92 ms, matrix 320 × 320 or 512 × 512, BW 390–413, ETL 20, NEX 2–3. To avoid that differences in imaging technique could cause different reliabilities in patients with and without disc prosthesis, both groups underwent the same metal-artifact reducing protocols.

The images were obtained in DICOM format and were de-identified before rating. CT and MRI examinations were mixed in random order, and the same patient's CT and MRI received different de-identification labels to ensure independent evaluations of CT and MRI findings.

Ratings

Blinded to clinical data, three experienced radiologists (> 15 years), two radiologists from the same institution with special interest in musculoskeletal radiology (A and B) and one neuroradiologist (C) from a different institution, independently interpreted the random mix of CT and MRI examinations in the same random order on a clinical Picture Archiving and Communication System (PACS) unit. FA was graded according to Weishaupt et al [16] as recommended in a review by Kettler and Wilke [22], and the observers were assigned example images from Weishaupt et al as support for decisions on grading [16]. Accordingly, the observers evaluated focal or general narrowing of the facet joint space (typically < 2 mm, recorded as present or not), osteophytes/hypertrophy (no/small/moderate/severe), erosions (no/mild/severe), and subchondral cysts (present or not) [16]. Based on these individual findings, a total grade of FA was reported (normal/mild/moderate/severe) (Figs. 1 and 2). Each finding and total grade FA were reported separately for L3/L4, L4/L5, and L5/S1 (worst side of right and left).

Pilot study

To synchronize their understanding of the rating criteria, the three observers independently rated CT and MRI of three patients (i.e., 9 levels/pairs of facets, 4 levels with disc prosthesis), MRI of three patients (9 levels without prosthesis), and CT of three patients (9 levels without prosthesis) from another study. Afterwards, they discussed ratings, discrepancies, and criteria.

Statistical analyses

All findings were dichotomized (see "Results"), and the prevalence of each category was calculated at each rated level for each observer. Prevalence of findings and kappa for pairwise interobserver agreement on total grade FA were calculated using SPSS statistics 24 (IBM). Using STATA 14 (StataCorp LLC), overall kappa was computed for agreement on total FA grade between all observers with a 95% bias-corrected confidence interval based on bootstrapping with 1000 repetitions.

Prevalence and bias (disagreement on prevalence) affect kappa, and "ordinary" kappa values are usually reported only for findings with a prevalence of 10–90%, since very low or high prevalence can lead to very low agreement beyond chance (i.e., low kappa values) despite very high actual agreement [23]. Accordingly, we calculated prevalence- and bias-adjusted kappa (PABAK) to compare interobserver agreement between different findings and different groups with potentially different or low prevalence of FA. We computed mean PABAK for total FA grade across all levels and observer pairs (mean of nine values) at CT and MRI and across L4/L5 and L5/S1 and all observer pairs (mean of six values) for levels with and levels without prosthesis (no prosthesis at L3/L4). We calculated the corresponding mean PABAK values for each individual finding. PABAK values were computed using

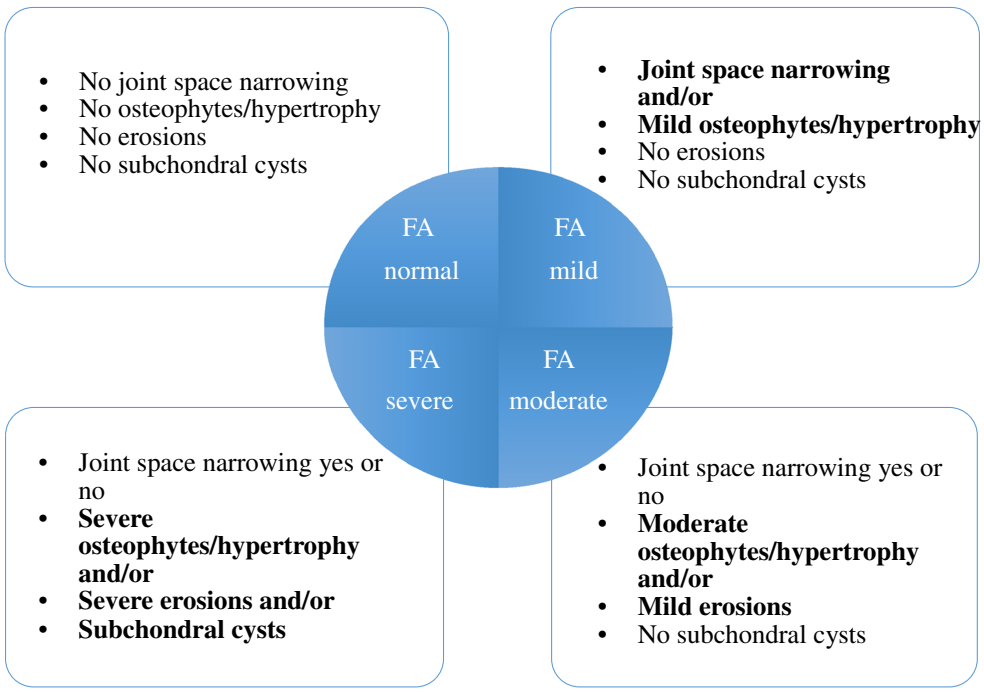

Fig. 1 Criteria for grading of facet arthropathy (FA) proposed by Weishaupt et al—the data in bold are criteria that distinguish the given FA grade from the previous one

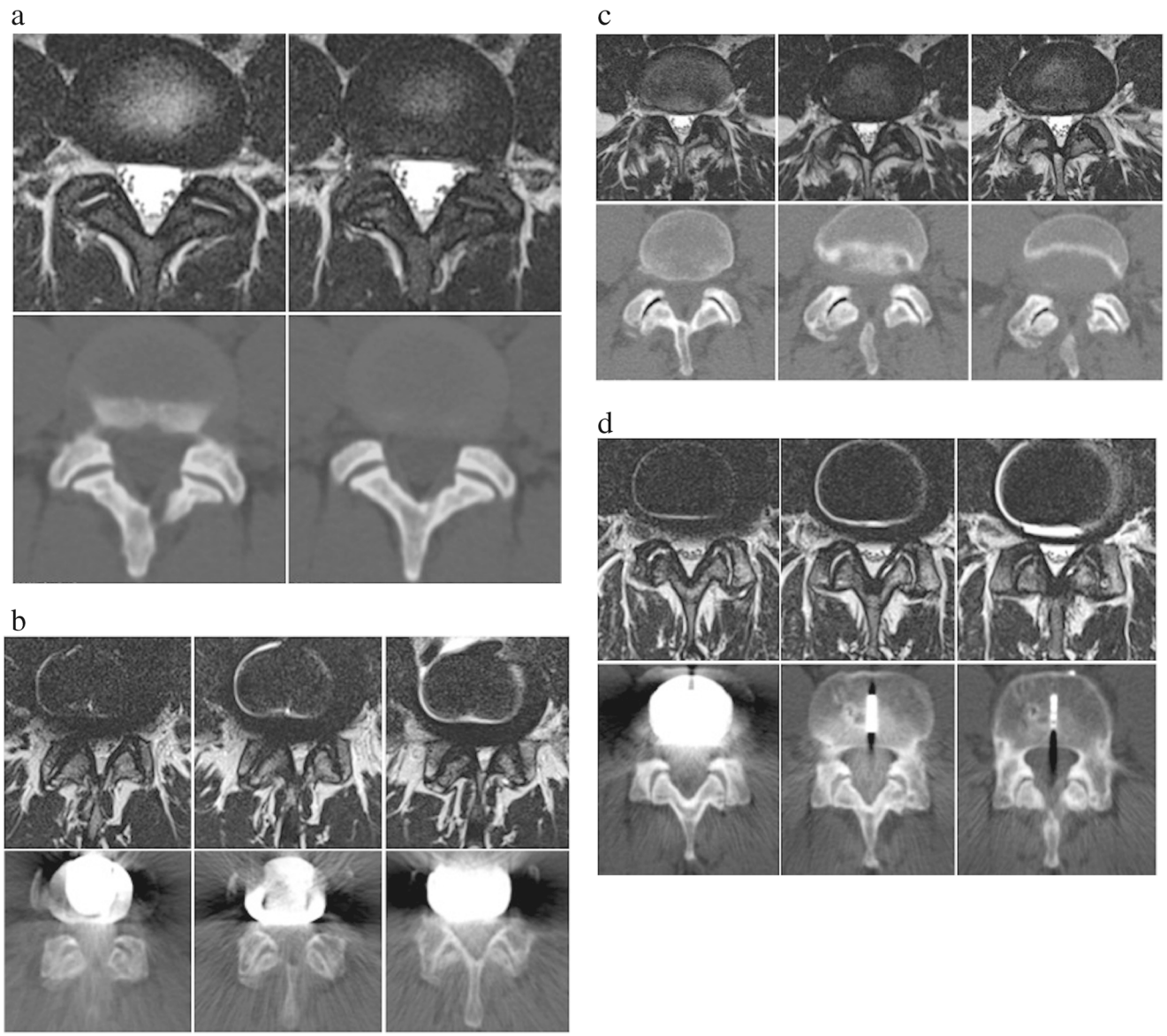

Fig. 2 Normal/mild (**a**), moderate (**b**), and severe (**c**, **d**) facet arthropathy (FA) on MRI and CT of patients without (**a**, **c**) and with (**b**, **d**) lumbar disc prosthesis. **a** 37-year-old woman treated non-operatively for chronic low back pain. Two consecutive axial MRI (T2, upper row) and CT images (lower row) at L4/L5 showing no FA according to one observer and mild FA according to two observers (disc space narrowing and/or mild osteophytes/hypertrophy on MRI and disc space narrowing on CT). (**b**) 53-year-old man treated with disc prosthesis for chronic low back pain. Three consecutive axial MRI (T2, upper row) and CT images (lower row) at L4/L5 showing moderate FA according to three observers (disc space narrowing and moderate osteophytes/hypertrophy both on MRI and CT—and, according to two observers, mild erosions on CT). (**c**) 55-year-old man treated non-operatively for chronic low back pain. Three consecutive axial MRI (T2, upper row) and CT images (lower row) at L4/L5 showing severe FA according to all three observers (disc space narrowing and severe osteophytes/hypertrophy on MRI and disc space narrowing, moderate osteophytes/hypertrophy, and mild erosions on CT—and, according to one observer each, mild erosions on MRI and subchondral cyst on CT). (**d**) 45-year-old man treated with disc prosthesis for chronic low back pain. Three consecutive axial MRI (T2, upper row) and CT images (lower row) at L4/L5 showing severe FA according to three observers (disc space narrowing, moderate osteophytes/hypertrophy, and subchondral cyst on MRI and CT—and mild erosions on CT according to two observers)

WinPepi 10.9 (http://www.brixtonhealth.com/pepi4windows.html) and were returned without confidence interval. We interpreted kappa and PABAK values as poor (≤0.20), fair (0.21–0.40), moderate (0.41–0.60), good (0.61–0.80), or very good (0.81–1.00) agreement beyond chance [24].

Results

The 114 included patients (54 men, 60 women) had mean age 49 years when imaged for this study; 66 patients (58%) had disc prosthesis at L4/L5 and/or L5/S1.

FA total grade dichotomized

The prevalence of moderate or severe total grade FA varied between observers at all levels and both at CT and MRI (Table 1). It varied up to threefold to fourfold at L3/L4 both on CT (10–33%) and on MRI (12–51%) and varied less at L4/L5 and L5/S1 (Table 1).

Overall agreement (kappa) on moderate or severe FA was moderate at CT (0.47–0.48) and poor to fair at MRI (0.20–0.32). Pairwise agreement (PABAK) was fair to good (0.33–0.78) at CT (Table 2) and poor to moderate (0.09–0.51) at MRI (Table 3). Mean PABAK for moderate or severe FA across all observer pairs and levels on CT was 0.53 and on MRI 0.31.

Individual findings dichotomized

Since the prevalence of many of the individual findings was not between 10 and 90%, PABAK was used to assess agreement on these findings. Pairwise interobserver agreement (PABAK) was similar for CT and MRI at all evaluated levels, except for osteophytes/hypertrophy where agreement was better for CT (0.33–0.74) (Table 2) than for MRI (0.08–0.58) (Table 3).

Mean PABAK across all levels and observer pairs indicated that agreement was very good for narrow joint space (mean PABAK 0.85 CT and 0.91 MRI), fair to moderate for osteophytes/hypertrophy (0.57 CT and 0.34 MRI), and moderate to good for erosions (0.78 CT and 0.81 MRI) and subchondral cysts (0.80 CT and 0.82 MRI). CT provided better agreement than MRI only for osteophytes/hypertrophy (mean PABAK 0.57 versus 0.34).

Impact of disc prosthesis

Mean PABAK across the two prosthesis levels and all observer pairs was similar at levels with and without prosthesis for all evaluations (difference ≤ 0.06 in both directions) (Table 4).

Discussion

Interobserver agreement on moderate or severe FA based on Weishaupt et al's grading system was moderate for CT (kappa 0.47–0.48, mean PABAK 0.54) and poor to fair for MRI (kappa 0.20–0.32, mean PABAK 0. 31). Agreement on facet osteophytes/hypertrophy was better for CT than MRI (mean PABAK 0.58 versus 0.35), but was poorer than agreement on other FA findings. Disc prosthesis at the rated level did not affect interobserver agreement.

To our knowledge, this study was the first to assess the reliability for each finding that the total FA grade is based on, exploring reasons for disagreement on FA. It was also the first to assess disagreement on prevalence of FA when using Weishaupt et al's recommended grading system. Only one prior study (published 1999) has compared the reliability of this system between CT and MRI [16], and only one study has made such a comparison for any other FA grading system

Table 1 Reported prevalence of findings in percent evaluated on CT and MRI ($n = 114$)

Finding, yes	Observer A		Observer C		Observer B	
	CT %	MRI %	CT %	MRI %	CT %	MRI %
L3/L4 FA total grade 2–3	23	32	33	51	10	12
Narrow joint space	92	91	83	98	98	100
Osteophytes/hypertrophy, moderate-severe	17	27	23	44	7	10
Erosions, mild-severe	2	5	10	11	0	3
Subchondral cyst	11	4	14	14	4	2
L4/L5 FA total grade 2–3	46	56	55	50	29	37
Narrow joint space	93	96	100	97	99	100
Osteophytes/hypertrophy, moderate-severe	36	54	42	42	24	34
Erosions, mild-severe	10	6	26	10	4	4
Subchondral cyst	14	5	19	17	9	4
L5/S1 FA total grade 2–3	61	64	46	36	34	44
Narrow joint space	99	95	94	97	98	99
Osteophytes/hypertrophy, moderate-severe	61	62	40	30	33	44
Erosions, mild-severe	4	5	10	5	2	3
Subchondral cyst	8	4	11	7	5	1

CT computed tomography, *MRI* magnetic resonance imaging, *FA* facet arthropathy, graded according to Weishaupt et al [6], *FA total grade 2–3* moderate-severe facet arthropathy

Table 2 Prevalence- and bias-adjusted kappa (PABAK) for individual observer pairs and overall kappa for interobserver agreement on findings on CT ($n = 114$)

Finding on CT, yes/no	Observers A/C		Observers A/B		Observers B/C		Overall kappa (95% CI)
	Disc prosthesis		Disc prosthesis		Disc prosthesis		
	No	Yes	No	Yes	No	Yes	
L3/L4 FA total grade 2–3	0.63	NA	0.74	NA	0.54	NA	0.47 (0.33–0.62)
Narrow joint space	0.68	NA	0.84	NA	0.67	NA	
Osteophyte/hypertrophy, moderate-severe	0.70	NA	0.74	NA	0.61	NA	
Erosions, mild-severe	0.81	NA	0.96	NA	0.81	NA	
Subchondral cyst	0.81	NA	0.84	NA	0.79	NA	
L4/L5 FA total grade 2–3	0.49	0.59	0.57	0.33	0.44	0.44	0.48 (0.36–0.61)
Narrow joint space	0.89	0.79	0.89	0.85	1.00	0.95	
Osteophyte/hypertrophy, moderate-severe	0.57	0.54	0.68	0.38	0.68	0.54	
Erosions, mild-severe	0.57	0.64	0.89	0.69	0.47	0.54	
Subchondral cyst	0.76	0.74	0.81	0.85	0.79	0.79	
L5/S1 FA total grade 2–3	0.36	0.78	0.38	0.41	0.56	0.63	0.47 (0.36–0.60)
Narrow joint space	0.91	0.93	1.00	1.00	0.91	0.93	
Osteophyte/hypertrophy, moderate-severe	0.33	0.70	0.36	0.41	0.61	0.70	
Erosions mild/severe	0.79	0.93	0.89	1.00	0.82	0.93	
Subchondral cyst	0.75	0.78	0.89	0.85	0.72	0.78	

CT computed tomography, *CI* confidence interval, *FA* facet arthropathy, graded according to Weishaupt et al [6], *NA* not applicable (no disc prosthesis at level L3/L4), *FA total grade 2–3* moderate-severe facet arthropathy

Table 3 Prevalence- and bias-adjusted kappa (PABAK) for individual observer pairs and overall kappa for interobserver agreement on findings on MRI ($n = 114$)

Finding on MRI, yes/no	Observers A/C		Observers A/B		Observers B/C		Overall kappa (95% CI)
	Disc prosthesis		Disc prosthesis		Disc prosthesis		
	No	Yes	No	Yes	No	Yes	
L3/L4 FA total grade 2–3	0.33	NA	0.51	NA	0.09	NA	0.20 (0.08–0.34)
Narrow joint space	0.82	NA	0.82	NA	0.96	NA	
Osteophyte/hypertrophy, moderate-severe	0.42	NA	0.58	NA	0.21	NA	
Erosions, mild-severe	0.74	NA	0.88	NA	0.75	NA	
Subchondral cyst	0.70	NA	0.91	NA	0.75	NA	
L4/L5 FA total grade 2–3	0.20	0.13	0.44	0.33	0.23	0.49	0.30 (0.17–0.41)
Narrow joint space	0.89	0.90	0.89	0.95	0.95	0.95	
Osteophyte/hypertrophy, moderate-severe	0.25	0.08	0.41	0.33	0.41	0.44	
Erosions, mild-severe	0.65	0.85	0.84	0.95	0.65	0.90	
Subchondral cyst	0.63	0.74	0.95	0.90	0.68	0.85	
L5/S1 FA total grade 2–3	0.24	0.19	0.45	0.48	0.29	0.26	0.32 [6]0.20–0.45)
Narrow joint space	0.86	0.93	0.89	0.93	0.93	1.00	
Osteophyte/hypertrophy, moderate-severe	0.20	0.26	0.47	0.41	0.36	0.26	
Erosions, mild-severe	0.84	0.78	0.91	0.78	0.89	0.70	
Subchondral cyst	0.82	0.78	0.93	0.93	0.89	0.85	

MRI magnetic resonance imaging, *CI* confidence interval, *FA* facet arthropathy, graded according to Weishaupt et al [6], *NA* not applicable (no disc prosthesis at level L3/L4), *FA total grade 2–3* moderate-severe facet arthropathy

Table 4 Mean prevalence- and bias-adjusted kappa (PABAK) for total FA grade and individual findings

Finding, yes/no	Disc prosthesis			
	No		Yes	
	CT	MRI	CT	MRI
FA total grade 2–3	0.47	0.31	0.53	0.31
Narrow joint space	0.93	0.90	0.91	0.94
Osteophyte/hypertrophy, moderate-severe	0.54	0.35	0.55	0.30
Erosions, mild-severe	0.74	0.80	0.79	0.83
Subchondral cyst	0.79	0.82	0.80	0.84

Each tabled value is the mean of six PABAK values from three observer pairs and two levels (L4/L5 and L5/S1)

CT computed tomography, *MRI* magnetic resonance imaging, *FA* facet arthropathy graded according to Weishaupt et al [6], *FA total grade 2–3* moderate-severe facet arthropathy

(that of Fujiwara et al [18], not recommended by Kettler and Wilke [22]). Our study was the first to apply metal-artifact reducing CT and MRI protocols when assessing the reliability of FA, improving the visibility near disc prostheses. FA can accelerate at the prosthesis level, but results for impact on outcome are conflicting [11, 12]. Clarification of impact of FA requires reliable FA grading.

The prevalence of moderate or severe FA varied between observers (particularly at L3/L4) both at CT and at MRI, mostly due to disagreement on moderate or severe versus mild or no osteophytes/hypertrophy (Table 1). Earlier data for disagreement on prevalence of FA based on Weishaupt's grading system are lacking, but data from Carrino et al support our finding [17].

The overall interobserver agreement on FA (kappa 0.47–0.48 for CT and 0.20–0.32 for MRI) was within the broad range of previous results. It should be noted, however, that it is difficult to compare kappa between studies because of differences in prevalence and in bias (which both affect kappa values), use of weighted versus unweighted kappa, different grading systems and dichotomization of the findings, different professions and experiences of the observers, and use of different or non-standardized CT and MRI techniques and equipment.

The prior study of Weishaupt's grading system reported higher but weighted kappa values for two musculoskeletal radiologists' non-dichotomized grading of FA both at CT (0.60) and at MRI (0.41) [16]. In a CT study of Weishaupt's system, kappa for interobserver agreement on FA was 0.59–0.94 (type of kappa not reported) [25]. In a MRI study of example-based grading of FA (normal, mild, moderate, severe), kappa for agreement between radiologists on normal/abnormal was 0.54 [17]. In studies of the different systems for grading FA by Fujiwara et al [26], kappa for interobserver agreement on FA was lower both for CT (0.27–0.33) [18] and for MRI (0.07–0.24) [15, 18].

We found better interobserver agreement on FA at CT than at MRI, mainly due to better agreement on osteophytes/hypertrophy. Weishaupt et al reported similar results for total FA grade (kappa 0.60 for CT and 0.41 for MRI) [16], but they found excellent agreement on FA at both CT and MRI if differences of one grade were disregarded.

As reported also in a study on change in MRI findings over time [27], disc prosthesis did not influence interobserver agreement on FA. With the CT and MRI protocols used in the present study, prosthesis artifacts are not an important source of variability in the assessment of FA.

Strengths and limitations

A main strength of our study was the estimation of reliability for each finding underlying the total FA grade. Further strengths were the metal-artifact reducing imaging protocols, the use of three experienced pre-trained radiologists as observers, and the rating of both CT and MRI examinations in random order (not CT first and MRI afterwards or vice versa) to avoid any learning effect during the study to bias the results for CT versus MRI. In addition, when comparing agreement between groups (e.g., levels with versus levels without prosthesis), we used PABAK to adjust for any between group differences in prevalence of FA findings.

Limitations were the fixed, moderate sample size and the low prevalence of some findings, implying few presentations of these findings to rate and a need to use PABAK in all analyses of agreement on these findings (and not only when comparing agreement between groups). PABAK can be useful, but reflects a hypothetical situation without any effect of prevalence and bias [23]. The PABAK values also lacked confidence interval and we could not compare them statistically. Furthermore, the observers could not be blinded to disc prosthesis (it was visible on the images), and how this may have affected their ratings is not clear.

We used a recommended system for grading FA [22], but other systems, FA findings, and imaging approaches are also relevant. The facet joints consist of cartilage, synovia, and capsule and may display similar degenerative findings as other true synovial joints: changes in cartilage, synovium, and capsule, followed by osteophytes or hypertrophy of bone, subchondral sclerosis, edema, and cysts [2, 16]. The two often used FA grading systems by Weishaupt [16] and Fujiwara [26] both include joint space narrowing and osteophyte. Weishaupt added subchondral erosions and cysts and Fujiwara added subchondral sclerosis. Other findings potentially relevant to pain (subchondral microfractures, synovitis, capsule distension, edema) are more easily seen on fat-suppressed MRI sequences, but are not part of the most commonly used grading systems [2]. A limitation of our study was that it did not include fat-suppressed MRI or other imaging that could have demonstrated such findings.

Importantly, the present results for reliability cannot indicate if or when FA is a clinically useful finding. They rather provide an improved basis for evaluating this issue. In summary, existing data on clinical relevance of structural FA are diverging, and this may partly be due to variable assessment of the finding. FA is often likely to be clinically irrelevant, but may also be relevant for symptoms, biomechanical function, or outcome in some patients. Reliable assessment of the finding is required before ascertaining its potential clinical relevance.

Implications

We suggest the following implications of our results. First, a more reliable grading of FA requires a more consistent evaluation of osteophytes/hypertrophy. Second, researchers should consider including CT and not only MRI in studies on FA, since CT can provide better reliability for total grade FA and osteophytes/hypertrophy. Third, metal-artifact reducing imaging protocols should be used when evaluating FA in patients with disc prosthesis, to improve visibility and maintain reliability at the prosthesis level. Fourth, one should further improve the reliability for the conclusive FA findings used in clinical research. One way to do this is to base the conclusive findings on multiple observers' ratings [28]. A computer-assisted reporting tool can also improve the reliability for FA [29]. Further optimized imaging technique might improve reliability too. Finally, our results underscore the importance of assessing the reliability of any FA finding before using it in research or clinical work.

In conclusion, interobserver agreement on FA was moderate for CT and better for CT than for MRI. Disc prosthesis at the rated level did not influence agreement. A more reliable grading of FA requires a more consistent evaluation of osteophytes/hypertrophy.

Acknowledgments Tone Berg Kjøndahl: punching data

Funding The authors state that this work has not received any funding.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Ansgar Espeland.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was obtained from all subjects (patients) in this study.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap Some study subjects or cohorts have been previously reported in Spine 2018 on adjacent level disc degeneration at 8-year follow-up of these RCT cohorts [30]. That publication contained no data on facet arthropathy. Other publications from the original RCT included baseline and 2-year data [19] and also reliability data [15, 27], but not any CT data and not any data on facet arthropathy evaluated using the same system as in the present manuscript. The most relevant earlier reports are referenced in the manuscript [4, 11, 19, 27].

Methodology
- retrospective
- cross-sectional study
- multicenter study

References

1. Goode AP, Carey TS, Jordan JM (2013) Low back pain and lumbar spine osteoarthritis: how are they related? Curr Rheumatol Rep 15:305
2. Gellhorn AC, Katz JN, Suri P (2013) Osteoarthritis of the spine: the facet joints. Nat Rev Rheumatol 9:216–224
3. Zhou X, Liu Y, Zhou S et al (2016) The correlation between radiographic and pathologic grading of lumbar facet joint degeneration. BMC Med Imaging 16:27
4. Berg L, Hellum C, Gjertsen Ø et al (2013) Do more MRI findings imply worse disability or more intense low back pain? A cross-sectional study of candidates for lumbar disc prosthesis. Skeletal Radiol 42:1593–1602
5. Endean A, Palmer KT, Coggon D (2011) Potential of magnetic resonance imaging findings to refine case definition for mechanical low back pain in epidemiological studies: a systematic review. Spine (Phila Pa 1976) 36:160–169
6. Varlotta GP, Lefkowitz TR, Schweitzer M et al (2011) The lumbar facet joint: a review of current knowledge: part 1: anatomy, biomechanics, and grading. Skeletal Radiol 40:13–23
7. Amoretti N, Iannessi A, Lesbats V et al (2012) Imaging of intervertebral disc prostheses. Diagn Interv Imaging 93:10–21
8. Murtagh RD, Quencer RM, Cohen DS, Yue JJ, Sklar EL (2009) Normal and abnormal imaging findings in lumbar total disk replacement: devices and complications. Radiographics 29:105–118
9. Salzmann SN, Plais N, Shue J, Girardi FP (2017) Lumbar disc replacement surgery-successes and obstacles to widespread adoption. Curr Rev Musculoskelet Med 10:153–159
10. Park CK, Ryu KS, Jee WH (2008) Degenerative changes of discs and facet joints in lumbar total disc replacement using ProDisc II: minimum two-year follow-up. Spine (Phila Pa 1976) 33:1755–1761
11. Hellum C, Berg L, Gjertsen Ø et al (2012) Adjacent level degeneration and facet arthropathy after disc prosthesis surgery or rehabilitation in patients with chronic low back pain and degenerative disc: second report of a randomized study. Spine (Phila Pa 1976) 37:2063–2073
12. Siepe CJ, Zelenkov P, Sauri-Barraza JC et al (2010) The fate of facet joint and adjacent level disc degeneration following total lumbar disc replacement: a prospective clinical, X-ray, and magnetic resonance imaging investigation. Spine (Phila Pa 1976) 35:1991–2003

13. Jarvik JG, Deyo RA (2009) Moderate versus mediocre: the reliability of spine MR data interpretations. Radiology 250:15–17
14. Feinstein AR (1983) An additional basic science for clinical medicine: IV. The development of clinimetrics. Ann Intern Med 99: 843–848
15. Berg L, Neckelmann G, Gjertsen Ø et al (2012) Reliability of MRI findings in candidates for lumbar disc prosthesis. Neuroradiology 54:699–707
16. Weishaupt D, Zanetti M, Boos N, Hodler J (1999) MR imaging and CT in osteoarthritis of the lumbar facet joints. Skeletal Radiol 28: 215–219
17. Carrino JA, Lurie JD, Tosteson AN et al (2009) Lumbar spine: reliability of MR imaging findings. Radiology 250:161–170
18. Stieber J, Quirno M, Cunningham M, Errico TJ, Bendo JA (2009) The reliability of computed tomography and magnetic resonance imaging grading of lumbar facet arthropathy in total disc replacement patients. Spine (Phila Pa 1976) 34:E833–E840
19. Hellum C, Johnsen LG, Storheim K et al (2011) Surgery with disc prosthesis versus rehabilitation in patients with low back pain and degenerative disc: two year follow-up of randomised study. BMJ 342:d2786
20. Stradiotti P, Curti A, Castellazzi G, Zerbi A (2009) Metal-related artifacts in instrumented spine. Techniques for reducing artifacts in CT and MRI: state of the art. Eur Spine J 18(Suppl 1):102–108
21. Marshman LA, Strong G, Trewhella M, Kasis A, Friesem T (2010) Minimizing ferromagnetic artefact with metallic lumbar total disc arthroplasty devices at adjacent segments: technical note. Spine (Phila Pa 1976) 35:252–256
22. Kettler A, Wilke HJ (2006) Review of existing grading systems for cervical or lumbar disc and facet joint degeneration. Eur Spine J 15: 705–718
23. Sim J, Wright CC (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 85: 257–268
24. Altman D (1991) Practical statistics for medical research. Chapman& Hall-CRC, New York, pp 396–439
25. Kalichman L, Li L, Kim DH et al (2008) Facet joint osteoarthritis and low back pain in the community-based population. Spine (Phila Pa 1976) 33:2560–2565
26. Fujiwara A, Tamai K, Yamato M et al (1999) The relationship between facet joint osteoarthritis and disc degeneration of the lumbar spine: an MRI study. Eur Spine J 8:396–401
27. Berg L, Gjertsen Ø, Hellum C et al (2012) Reliability of change in lumbar MRI findings over time in patients with and without disc prosthesis–comparing two different image evaluation methods. Skeletal Radiol 41:1547–1557
28. Espeland A, Vetti N, Kråkenes J (2013) Are two readers more reliable than one? A study of upper neck ligament scoring on magnetic resonance images. BMC Med Imaging 13:4
29. Wang B, Rosenthal DI, Xu C et al (2018) The effect of computer-assisted reporting on interreader variability of lumbar spine MRI degenerative findings: five readers with 30 disc levels. J Am Coll Radiol. https://doi.org/10.1016/j.jacr.2017.12.020
30. Furunes H, Hellum C, Espeland A et al (2018) Adjacent disc degeneration after lumbar total disc replacement or non-operative treatment: a randomized study with 8-year follow-up. Spine (Phila Pa 1976). https://doi.org/10.1097/BRS.0000000000002712