Osteoarthritis and Cartilage 22 (2014) 1360–1366

Osteoarthritis and Cartilage

OARSI OSTEOARTHRITIS RESEARCH SOCIETY INTERNATIONAL

A new CT grading system for hip osteoarthritis

T.D. Turmezei †‡§*, A. Fotiadou ||, D.J. Lomas ‡, M.A. Hopper ‡, K.E.S. Poole §

† Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK
‡ Department of Radiology, Box 218, Level 5, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK
§ Department of Medicine, Box 157, Level 5, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK
|| Department of Radiology, Hinchingbrooke Health Care NHS Trust, Hinchingbrooke Hospital, Hinchingbrooke Park, Huntingdon PE29 6NT, UK

ARTICLE INFO

Article history:
Received 6 January 2014
Accepted 4 March 2014

Keywords:
Osteoarthritis
Hip joint
Computed tomography
Grading
Reliability

SUMMARY

Objectives: We have developed a new grading system for hip osteoarthritis using clinical computed tomography (CT). This technique was compared with Kellgren and Lawrence (K&L) grading and minimum joint space width (JSW) measurement in digitally reconstructed radiographs (DRRs) from the same CT data. In this paper we evaluate and compare the accuracy and reliability of these measures in the assessment of radiological disease.

Design: CT imaging of hips from 30 female volunteers aged 66 ± 17 years were used in two reproducibility studies, one testing the reliability of the new system, the other testing K&L grading and minimum JSW measurement in DRRs.

Results: Intra- and inter-observer reliability was substantial for CT grading according to weighted kappa (0.74 and 0.75 respectively), while intra- and inter-observer reliability was at worst moderate (0.57) and substantial (0.63) respectively for DRR K&L grading. Bland–Altman analysis showed a systematic difference in minimum JSW measurement of 0.82 mm between reviewers, with a least detectable difference of 1.06 mm. The area under the curve from ROC analysis was 0.91 for our CT composite score.

Conclusions: CT grading of hip osteoarthritis (categorised as none, developing and established) has substantial reliability. Sensitivity was increased when CT features of osteoarthritis were assigned a composite score (0 = none to 7 = severest) that also performed well as a diagnostic test, but at the cost of reliability. Having established feasibility and reliability for this new CT system, sensitivity testing and validation against clinical measures of hip osteoarthritis will now be performed.

Introduction

Two important and unmet challenges in osteoarthritis imaging are the detection of clinically relevant early disease and accurate prediction of disease progression. While some tissue biomarkers have shown promise in relation to prediction of disease progression at the knee and hip[1,2], imaging has the advantage of representing disease at specific locations around the body. The relationship between imaging features and symptomatic disease is also beginning to emerge, even if the relationship between pain and structural disease remains obscure[3].The presence of bone marrow lesions detected with magnetic resonance imaging (MRI) at specific locations in the knee for individuals without symptoms has been shown to predict the onset of symptoms at 15 months, reinforcing the value of imaging pre-radiographic disease[4]. There is also scope for developing imaging and biochemical markers for osteoarthritis in parallel, since early metabolic disturbances in joint tissues are also linked with the later appearance of imaging lesions[5].

It is key to identify valid biomarkers that reflect clinically relevant changes in disease that can be used to (1) monitor the efficacy of new therapies in a trial setting, and (2) predict individuals at risk of new or rapidly deteriorating disease. Radiography and MRI have been the techniques most widely applied to hip osteoarthritis. However, there is a balance between the capabilities of MRI and radiographic evaluation in respect of what features they can visualise and how sensitively they do it. In our first paper on the CT assessment of hip osteoarthritis[35], we characterised computed tomography (CT) features of hip osteoarthritis according to location and severity, which has led to the development of the CT grading system presented here. This grading system relies on the 3D interpretation of bone-related imaging features around the femoral

* Address correspondence and reprint requests to: T.D. Turmezei, Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK.
E-mail address: tom@diagnosticradiology.eu (T.D. Turmezei).

http://dx.doi.org/10.1016/j.joca.2014.03.008
1063-4584/

head, specifically osteophytes, subchondral cysts and minimum joint space width (JSW). We believe that CT has significant advantages over MRI and radiographs for visualising these features, and offers a means to stratify and phenotype disease more accurately than the current standard of radiography. This opinion is based on previous arguments that have considered how CT is not only excellent for visualisation of such features, but also how it may enhance our understanding of disease[6,7].

Following on from our description of feature severity mapping of hip osteoarthritis with unenhanced clinical CT[35], in this paper we introduce the construct of a new CT grading system of hip osteoarthritis and compare its reliability with Kellgren & Lawrence (K&L) grading and minimum JSW measurement in digitally reconstructed radiographs (DRRs) from the same CT data. We also show how these imaging scores correlate in the assessment of radiological disease and report the performance of CT as a diagnostic test, both important steps in developing CT as an imaging biomarker of hip osteoarthritis.

Methods

This study involved the same cohort of 247 female volunteers who had consented to the use of their CT imaging for the investigation of hip disease as detailed in the first paper in this series by Turmezei *et al.*[35], with application of the same exclusion criteria. Forty individuals were selected from the cohort by the first author (TT) during a preliminary review of the axial CT imaging to include a range of osteoarthritis imaging features from absent to severe. Plain radiographic imaging and clinical scores of hip disease were not available for this study. Participant mean age was 66 ± 17 years standard deviation (SD), ranging from 27 to 90 years (Table I).

Imaging of both hips from each individual was included in the two reproducibility studies, yielding a total of 80 hips for assessment in each. All imaging was without exogenous contrast medium and acquired on clinical CT scanners with slice thickness ranging from 0.75 to 1.5 mm; full clinical CT scanner and acquisition details are given in Turmezei *et al.*[35]. Each individual's imaging was fully anonymised and given a unique study identifier. Ten individuals (20 hips) were randomly selected as test cases for practice assessment. The remaining 30 individuals (60 hips) were used for the reproducibility studies. Each set of hips remained paired for imaging review.

Prior to each reproducibility study, reviewers had a 1-h meeting with the study organiser (TT) for discussion of study protocol and methodology and were then free to examine test cases over the next few weeks. A second meeting followed this familiarisation period to cover any questions on methodology. The first interpretation run was performed in a randomised order. Each individual's imaging was re-randomised and ascribed a new identifier for the second interpretation run, which was performed at least 4 weeks after the first. TT (a radiologist who has completed training with musculoskeletal subspecialisation) and DL (a professor of radiology with musculoskeletal imaging expertise) were reviewers for the CT scoring/grading study. TT and AF (a consultant musculoskeletal radiologist) were reviewers for the DRR study. The same sets of imaging were used for both reproducibility studies. The DRR reproducibility study was performed 1 year after the CT grading study.

Table I
Demographics from the 40 cases selected for the reproducibility studies (REPRO), a subset of the 230 individuals from the whole cohort (WHOLE) with imaging available for both hips and after exclusion criteria were applied

	n	Age ± SD (yr) [Range]	Weight ± SD (kg)	Height ± SD (m)	Body mass index ± SD (kg/m^2)
REPRO	40	66 ± 17 [27–90]	67.4 ± 11.5*	1.62 ± 0.08*	25.6 ± 4.2
WHOLE	230	66 ± 17 [27–98]	69.3 ± 14.2†	1.61 ± 0.08†	26.6 ± 5.3

* For four individuals, height and weight data was not recorded and so primary care records were used to provide a measurement from as near to the time of imaging as possible. For one of these, no height data was available, and so the patient was ascribed the mean value.

† For an additional 11 individuals, height and weight data was not recorded and so the primary care records were used to provide a measurement from as near to the time of imaging as possible.

CT grading reproducibility study

All imaging was reviewed on a clinical workstation equipped with GE AW Volume Share 2 software, AW version 4.4 (GE Healthcare, Milwaukee, Wisconsin, USA: www.gehealthcare.com/euen/advantage-workstation/index.html) with the fixed window level (2000 HU) and width (350 HU) and a magnification of up to 200%. Features were scored around each hip using the same multiplanar reformat approach as described for feature severity mapping by Turmezei *et al.*[35].

Completed scoresheets were interpreted at the end of the study in accordance with the system given in Table II. This yielded a new CT-based score for each feature from around the femoral head: osteophytes (0–3), cysts (0–1), and JSW (0–3). See SI Fig. 1 and Turmezei *et al.*[35] for a description of assessment exclusion zones. Individual feature scores were combined as a CT composite score (0–7), 0 being the least and 7 the most severe representation of radiological osteoarthritis features. CT composite score was further divided into three broader CT grades: 0–2 as no radiological osteoarthritis; 3–4 as developing radiological osteoarthritis; 5–7 as established radiological osteoarthritis (Table II).

DRR study

All imaging was reviewed on an iMac (©2013 Apple Inc; 2.8 GHz Intel core i7, 4 GB RAM, AMD Radeon HD 6770M 512 MB graphic,

Table II
Interpretation of feature severity mapping scoresheets described in Turmezei *et al.*[35] used to generate individual feature scores summated for the CT composite score. The CT composite score can then be broken down into three separate CT grades

Osteophyte score (excluding the reaction area & fovea)	
0	0–4 (sum of osteophyte sector scores derived from severity mapping)*
1	5–9
2	10–19
3	>20
Subchondral cyst score (excluding the neck or pit area)	
0	Everything but grade 1 (grade derived from severity mapping)*
1	Any grade 3
JSW score (number of sectors with score 3, i.e., JSW <1.5 mm)	
0	0–1 sector (derived from severity mapping)*
1	2–3 sectors
2	4–5 sectors
3	6–7 sectors

CT composite score†	CT grade	Verbal interpretation of CT grade
0–2	None (0)	No radiological osteoarthritis
3–4	Developing (1)	Developing radiological osteoarthritis
5–7	Established (2)	Established features of radiological osteoarthritis

* See Turmezei *et al.*[35].

† CT composite score (0–7) obtained by summing osteophyte, subchondral cyst and JSW contigents.

1920 × 1080 display, Mac OS X Lion 10.7.3, Apple, Cupertino, http://www.apple.com/) using OsiriX DICOM viewer software (©Pixmeo Sarl; v.3.9.3 32-bit, http://www.pixmeo.com). Multiplanar reformatting was used to create a mean intensity projection slab aligned symmetrically in the axial plane to a true sagittal plane through the pubic symphysis, with coronal coverage of the anterior and posterior hip joint margins [Fig. 1(a)]. This resulted in a slab thickness from approximately 6 to 8 cm in coronal depth that could be reviewed in the coronal plane [Fig. 1(b)], a DRR surrogate of a true anteroposterior (AP) pelvic radiograph. A standard window level (200 HU) and window width (700 HU) was chosen to display the image with the estimated brightness and contrast of such a radiograph. A magnification of up to 200% was allowed for measurement of minimum JSW using the software electronic calliper tool and for K&L grading as defined in Table III.

Statistical analysis

A weighted kappa statistic was calculated for intra- and inter-observer agreement for discrete categorical data (i.e., individual feature scores, CT composite score, CT grade and K&L grading). This was preferred to the unweighted kappa statistic because it takes into account the degree of disagreement rather than relying solely on agreement. Kendall's tau rank correlation coefficient was calculated to assess correlation between discrete categorical data (i.e., CT composite score, CT grade and K&L grading) and continuous data (minimum JSW), and between discrete categorical data of different scales (e.g., CT composite score or CT grade vs K&L grade). Both were calculated using R v2.15.1 [R.app GUI 1.52 (6188 Leopard build 32-bit), S. Urbanek & H.-J. Bibiko, ©R Foundation for Statistical Computing, 2012]. The confidence interval (CI) for the weighted kappa statistic was calculated as ±1.96 times by the standard error. Bland–Altman plots were used to compare intra- and inter-observer reliability for continuous data (minimum JSW) using Microsoft Excel for Mac 2011, v14.2.3 (©2010 Microsoft Corporation). These plots yielded bias, limits of agreement and coefficient of variability, also termed least detectable difference. The coefficient of variability was calculated as ±1.96 times by the SD of the difference in first and second measurement. Receiver operating characteristic (ROC) graphs and statistics were calculated for CT composite score and CT osteophyte score for TT and DL using "ROC Analysis: web-based calculator for ROC curves", Johns Hopkins University, Baltimore, Maryland, USA (http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html). DRR K&L grading performed by AF was used as the gold standard for radiological diagnosis of hip osteoarthritis with a threshold of ≥2.

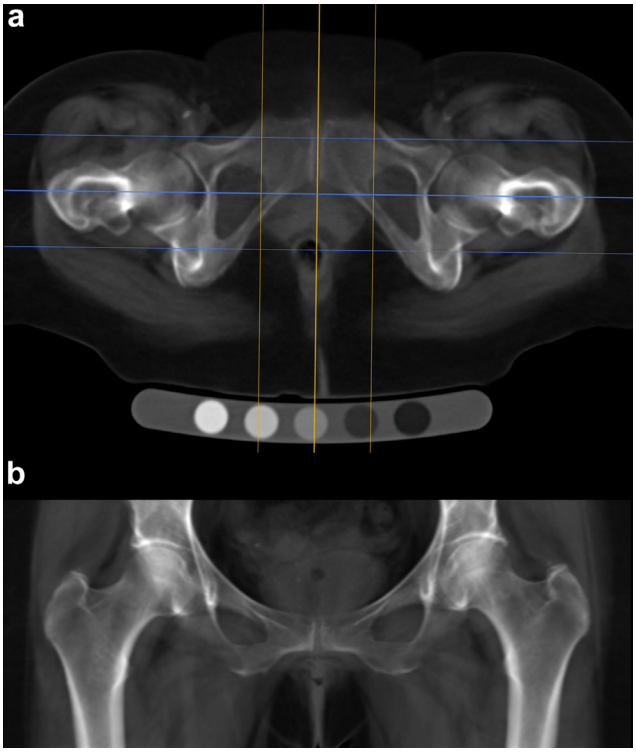

Fig. 1. DRR process from helically acquired CT data using OsiriX. (a) Axial mean intensity projection reformat of the original data showing the sagittal (orange) plane used to align along the AP axis of the pubic symphysis and the coronal (cyan) reformat plane with outer lines marking the limits of the reconstructed slab just beyond the anterior and posterior hip joint margins. (b) Coronal mean intensity projection slab (usually 6–8 cm in depth) showing the DRR used for minimum JSW measurement and K&L grading (window level 200; window width 700; magnification up to 200%).

Table III
Interpretation of the K&L grading system applied to DRRs

K&L score	Verbal grade	Feature description
0	None	No features of OA
1	Possible	Possible osteophytes or possible joint space narrowing (JSN)
2	Mild	Definite osteophytes **and/or** definite JSN
3	Moderate	Definite osteophytes, definite JSN, sclerosis, cysts, possible deformity
4	Severe	Grade 3 **plus** definite deformity (with or without severe sclerosis)

Results

Kappa statistics

Weighted kappa statistics for intra- and inter-observer reliability for individual features (osteophytes, subchondral cysts and JSW scores), CT composite score, CT grade and DRR K&L grade are presented in Table IV. Weighted kappa statistics for individual feature scores showed that CT scoring of osteophytes was substantial to near perfect for intra-observer reliability (0.78 and 0.87) and substantial for inter-observer reliability (0.62). Cyst scores had perfect to near perfect reliability (0.85–1.00). JSW scores showed substantial intra-observer reliability for TT (0.63), but only fair intra-observer reliability for DL (0.23) and inter-observer reliability (0.28).

Combining individual features as a CT composite score (as generated in Table II), reliability was substantial for intra-observer rating (0.65 and 0.64), and moderate for inter-observer rating (0.58). CT grading yielded uniformly substantial reliability (0.74–0.75). The reliability for DRR K&L grading was less consistent, being near perfect for TT's intra-observer reliability (0.84) and moderate for inter-observer reliability (0.63).

Bland–Altman plots

TT's intra-observer bias (limits of agreement) was 0.12 mm (−0.64–0.88 mm), meaning that there was a systematic overestimate of 0.12 mm per measurement on the first run compared to the second run, with a least detectable difference of 0.76 mm. AF's intra-observer bias was 0.10 mm (−0.63–0.83 mm) with a least detectable difference of 0.73 mm. Inter-observer bias was 0.82 mm (−0.25–1.88 mm) with a least detectable difference of 1.06 mm. Bland–Altman plots for JSW measurement are presented in Fig. 2.

Table IV
Individual feature scores, composite CT score, CT grading and DRR K&L grading reliability ratings (Key for kappa statistic agreement is also given*.)

	Intra-observer weighted kappa statistic (95% CI) (TT vs self)	Intra-observer weighted kappa statistic (95% CI) (other vs self)	Inter-observer weighted kappa statistic (95% CI) (TT vs other)
Osteophyte score	0.78 (0.51–1.00)	0.87 (0.68–1.00)	0.62 (0.39–0.86)
Cyst score	1.00 (0.91–1.00)	0.85 (0.75–0.94)	1.00 (0.91–1.00)
JSW score	0.63 (0.35–0.90)	0.23 (0.00–0.69)	0.28 (0.00–0.62)
CT composite score	0.65 (0.36–0.94)	0.64 (0.37–0.91)	0.58 (0.29–0.87)
CT grade	0.74 (0.47–1.00)	0.74 (0.53–0.95)	0.75 (0.48–1.00)
DRR K&L grading	0.84 (0.57–1.00)	0.57 (0.40–0.74)	0.63 (0.37–0.90)

**Kappa statistic agreement*: 0.81–1.00 = almost perfect; 0.61–0.80 = substantial; 0.41–0.60 = moderate; 0.21–0.40 = fair; 0.00–0.20 = slight; <0 = none.

Kendall's tau correlation coefficient

Kendall's tau correlation coefficients were calculated for each of: CT composite score vs K&L grade (0.566, $P < 0.001$); CT grade vs K&L grade (0.555, $P < 0.001$); CT composite score vs minimum JSW (−0.542, $P < 0.001$); CT grade vs minimum JSW (−0.509, $P < 0.001$); CT composite score vs CT grade (0.755, $P < 0.001$); and K&L grade vs minimum JSW (−0.509, $P < 0.001$). Note that minimum JSW is negatively correlated with other disease measures because lower JSW values are associated with worse disease.

ROC graphs and statistics

The graphs for ROC analysis of CT composite score and CT osteophyte score for TT and DL are shown in Fig. 3. Sensitivity and specificity for CT score were 80% and 82% for TT respectively and 70% and 86% for DL respectively. CT score accuracy was 82% for TT with a ROC area under the curve (AUC) of 0.91. CT score accuracy was 83% for DL with an AUC of 0.81. Sensitivity and specificity for CT osteophyte score were 50% and 90% for TT respectively and 40% and 96% for DL respectively. CT osteophyte score accuracy was 83% for TT with an AUC of 0.82. CT osteophyte score accuracy was 87% for DL with an AUC of 0.83. These results are shown in Table V.

Discussion

In this second paper on the unenhanced CT assessment of hip osteoarthritis, we have presented reproducibility statistics for new measures of radiological hip osteoarthritis, namely individual feature CT scores, CT composite score and CT grade, which have been derived from our newly described technique of feature severity mapping. We have presented these alongside two established measures (K&L grade and minimum JSW) applied to digital radiographs reconstructed from the same CT data. We have also assessed the performance of CT composite score and osteophyte score as a diagnostic test for disease against the gold standard of K&L grading. Our motivation has been to establish the reliability and accuracy of a new CT-based approach to the imaging assessment of radiological hip osteoarthritis against current radiographic standards. Since comparative radiographic imaging was not available for this study, DRRs were created as a surrogate for the assessment of radiographic disease. Although we applied a simple method of coronal plane digital radiograph reconstruction to axial CT data that did not take into account factors such as scatter and beam hardening[8], we have nonetheless demonstrated the feasibility of yielding such information from CT imaging, allowing it to serve as a benchmark for these CT measures.

The substantial to almost perfect reliability of femoral osteophyte scoring was an interesting result, with osteophytes being a consistently reliable feature across reviewers. This warrants further investigation to determine how osteophytes are related to clinically relevant disease, especially since Arden *et al.* recommended they be considered along with radiographic JSW for the detection of incident hip osteoarthritis[9]. The individual osteophyte feature score of

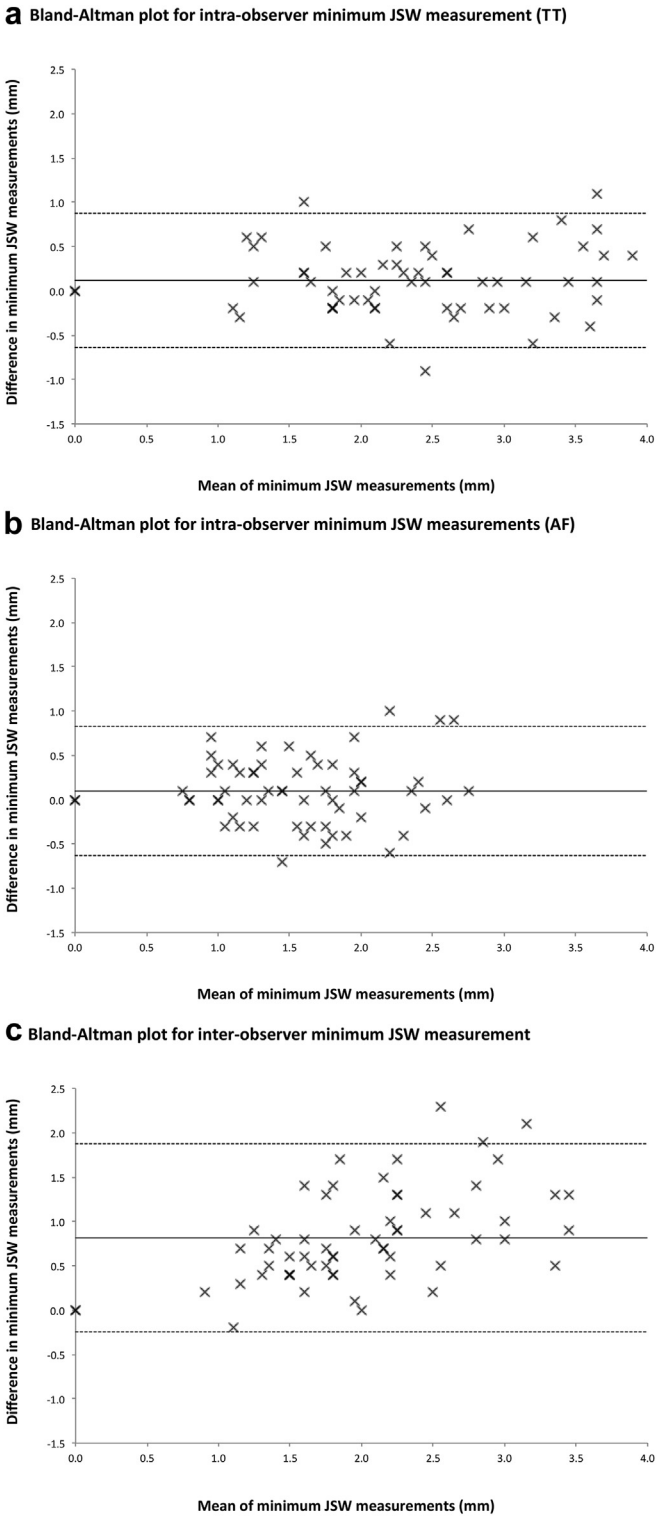

Fig. 2. Bland–Altman plots for intra- and inter-observer minimum JSW measurement.

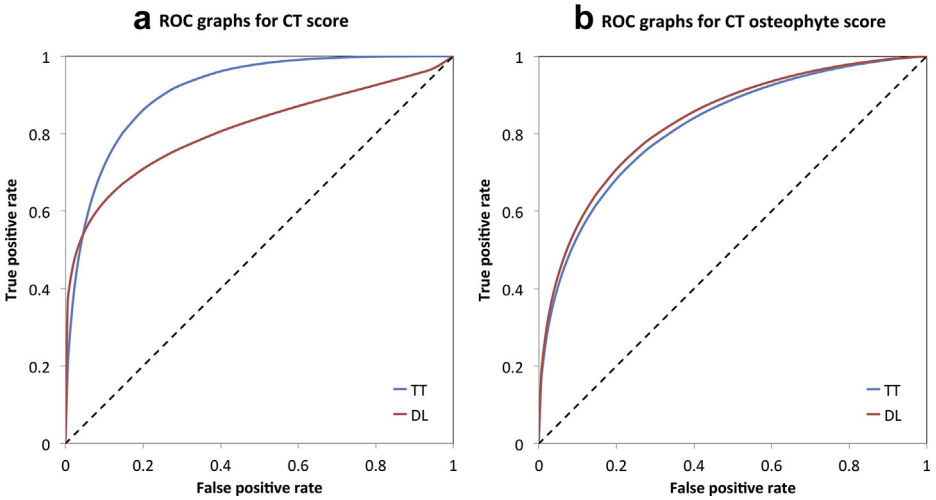

Fig. 3. ROC analysis graphs for (a) CT composite score and (b) CT osteophyte score comparing reviewer performance.

0–3 is derived from a much wider score range that represents overall osteophyte load from around the femur. This could theoretically reach 126 if each of the 42 sectors contained a severe osteophyte, however this is unlikely because femoral head subarticular osteophytes were rarely seen in the cohort, reflected by the highest osteophyte load score being 47 *ex* 126 and visually correlated with extremely severe osteophytosis around the femoral head and neck. However, the value of a wider osteophyte score range is that it could introduce greater sensitivity to the follow-up of disease.

As a condensation of the CT composite score, CT grading would mathematically have greater reliability, as demonstrated. The important consideration here is now in trade-off: more categories (eight for CT composite score compared to three for CT grade) will allow for greater sensitivity in detecting disease changes, but might result in poorer reliability. Nonetheless, the consistent substantial reliability for CT grade is support for taking this system forward to validation studies involving correlation with clinical outcome measures. This work is now in progress.

We cannot ignore the fact that radiographs have endured as the mainstay for clinical and research assessment of large joint osteoarthritis since the 1950s[10–12]. Historically K&L grading has been most popular, although it has evolved into multiple interpretations. For example, grade two knee osteoarthritis has been described in at least five different ways with different threshold criteria for disease[13,14]. This is, in part, because K&L did not provide a categorical description of disease features in their original 1957 paper. In fact this was first described in a report led by Kellgren in 1963 on a symposium entitled '*The Epidemiology of Chronic Rheumatism*' organised by The Council for the International Organizations of Medical Sciences[15,16]. We showed reliability of DRR K&L grading to be variable, perhaps exposing the suspected weakness of how to interpret its verbal instructions into a categorical score. It must also be considered that independent centres may be applying different interpretations of the grading. Significant correlation between scores of disease according to Kendall's tau correlation coefficient nonetheless indicates that the different assessment scores investigated in this study must be measuring a similar imaging manifestation of hip osteoarthritis.

Table V
ROC analysis of reviewer performance with CT score and osteophyte score using K&L grading from DRRs performed by a separate radiologist as the gold standard

Reviewer	Method	Sensitivity	Specificity	Accuracy	AUC*
TT	CT composite score	80%	82%	82%	0.91
	CT osteophyte score	50%	90%	83%	0.82
DL	CT composite score	70%	86%	83%	0.81
	CT osteophyte score	40%	96%	87%	0.83

* *AUC interpretation*: 0.90–1.00 = excellent; 0.81–0.90 = good; 0.71–0.80 = fair; 0.61–0.70 = poor; 0.50–0.60 = fail.

Inconsistent application of radiographic grading has been one motivation for JSW becoming an important radiographic measure of disease progression, as recommended by the official bodies of OARSI (Osteoarthritis Research Society International) and OMERACT (Outcome Measures in Rheumatology) in 2009[17]. These organisations performed a systematic literature review followed by an expert opinion review in order to reach a conclusion on the definition of relevant radiological progression in hip and knee osteoarthritis. No absolute cut-off for relevant JSW was determined at either joint, instead recommending that…

> "…*a cut-off should be determined for each study based on a pilot study that assesses the inherent variability of the measurement process in a representative sample of the studied population*"[17].

Our CT electronic calliper JSW measurement technique performed poorly in reproducibility analysis, most likely because of the repeated variability of relying on imaging planes set for each review, each time, for each case, by each assessor. One systematic review concluded that radiographic hip JSW had a weak association with clinical symptoms, yet still had predictive validity for future total hip replacement[18]. Another study recommended that researchers consider femoral osteophytes and JSW in composite definitions of disease for the best representation of incident radiographic hip osteoarthritis[9]. These are both features that can be assessed in 3D with CT, as we demonstrated with our paper on the assessment of hip osteoarthritis with feature severity mapping[35]. In our hands Bland–Altman analysis revealed that minimum JSW measured in DRRs was a reliable measure for individual reviewers but, like CT-based JSW measurement, was unreliable between reviewers with a systematic discrepancy of 0.82 mm and a least detectable difference of 1.06 mm. The least detectable difference for individual reviewers was 0.76 mm (TT) and 0.82 mm (AF), suggesting that any observed difference in DRR minimum JSW measurement would have to be at least these values before it could be considered real. Given that the largest minimum JSW measurement by TT was 4.0 mm, the percentage error in repeated performance would at best be 18%, and worse for smaller measurements.

This also brings into consideration the accuracy of electronic calliper measurement. DICOM imaging software electronic callipers have a geometric accuracy that depends on the ratio of the display field of view (DFOV) and the pixel matrix size. For the AW software, the DFOV was 13.0 cm with a pixel matrix of 512 × 512, providing a lower bound of geometric accuracy for length measurement of ±0.25 mm. This offers further explanation for the variable reliability for CT-based JSW scores on top of MPR positional factors. For OsiriX software electronic callipers (length ROI), DFOV was 622 × 280 mm with a pixel matrix of 1846 × 822, providing a lower bound of geometric accuracy for length measurement of ±0.34 mm. If we again consider the largest DRR minimum JSW measurement of 4.0 mm, this introduces a best potential error for single measurement of ±8.5%. Important as such technical sources of error may be, they are insurmountable with manual measurement, inviting the application of accurate and precise automated techniques for JSW measurement similar to those that have already been applied to cortical bone thickness in the setting of osteoporosis and fracture risk[19,20]. In fact, significant results have already been reported from cortical thickness mapping relating thicker peri-articular cortical bone and subchondral bone plate to worsening imaging features of hip osteoarthritis[21]. Advances in image analysis leading to automated 3D JSW measurement would remove such operator dependencies, which is the focus of our on-going research.

If these new CT measures of disease are to be considered as a diagnostic biomarker, it is essential to analyse their ROC performance. We recognise that we did not have clinical measures in this study to use as a gold standard to compare CT composite score and osteophyte score with K&L grading. This is the next phase of our research. In lieu of this, we were able to compare CT composite score and osteophyte score from two of our reviewers (TT and DL) against K&L grading from the third (AF). These results showed that CT composite score may perform well as a diagnostic test (with AUCs representing good to excellent performance), but they also showed a very high specificity for CT osteophyte score that was combined with a low sensitivity. This again suggests that the presence of osteophytes is an important consideration for disease detection and assessment[9]. A further interpretation of this result could be that achieving the diagnostic threshold of a K&L grade ≥2 relies substantially on osteophytes as an imaging feature, yet they are not always present with radiologically or clinically confirmed disease, as seen in patterns that have been described as 'atrophic'[22,23]. Thus CT osteophyte score is set to be specific for disease as categorised by K&L grade. Both systems now need to be compared against the clinical gold standard.

In any case, the possibility of imaging hip osteoarthritis with CT presents several opportunities that should compliment rather than supplant MRI. The performance of quantitative cartilage imaging with MRI of characteristics such as cartilage thickness, volume, T1-rho, and T2 values is compelling[24–29], as is its ability to detect significant bone-related pre-radiographic findings[4]. CT cannot compete directly with such measures, but it can offer a detailed assessment of mineralised bone including osteophytes[30], subchondral cysts[31], subchondral bone plate[21] and trabecular bone density[32]. CT can also provide a 3D perspective that has traditionally been provided in 2D by radiographs. This is particulary salient given the recognition that bone shape now has in the aetiology of hip osteoarthritis as femoro-acetabular impingement in particular[33]. Assessment of 3D shape as a risk factor for hip osteoarthritis is yet to be performed, but CT would be a strong candidate modality with which to proceed. Therefore, given the value of disease stratification for clinical trials[34], it is important to consider what CT may bring to the assessment of structural disease, especially since it can be faster, cheaper and more available than MRI. Indeed, a study comparing and combining MRI and CT in the assessment of large joint osteoarthritis would be an interesting and important undertaking.

Conclusion

Having reported on the ability of clinical CT to assess imaging features of hip osteoarthritis in the first of two papers[35], we have now constructed and tested the reliability of CT-based assessment. CT grading (none, developing, established) showed substantial reproducibility, with the potential to increase sensitivity by using a CT composite score (0–7), albeit at the cost of reduced reliability. CT-detected femoral osteophytes have also shown excellent reproducibility and specificity with our method and warrant further consideration as a reliable marker of disease. CT composite score and grade also correlated with traditional measures of K&L grade and minimum JSW made in DRRs from the same individuals, with CT score performing well as a diagnostic test compared to K&L grading.

Several steps will now follow on from this study, namely testing new CT score sensitivity to change with follow-up imaging and validation through correlation with clinical disease, cortical bone mapping for quantitative 3D subchondral bone assessment in the same cohort[21], and developing means of automated 3D hip joint space representation from clinical CT data.

Contributions

TT, DL, MH and KP contributed to conception and design of this study.

TT, AF, DL and MH performed data collection.

TT, DL and KP conducted data analysis.

TT, DL, and KP contributed to data interpretation and preparation of the manuscript.

The final version of the article was approved by all the authors.

TT takes responsibility for the integrity of the work as a whole.

Conflict of interests

None declared.

Funding

KP acknowledges support of an Arthritis Research UK Research Progression award, and the Cambridge NIHR Biomedical Research Centre (MEBB Theme). TT acknowledges the support of an Evelyn Trust Clinical Training Fellowship award. None of the funding sources had a role in study design, data handling, writing of the report, or decision to submit the paper for publication.

Acknowledgements

All the authors acknowledge the expert statistical guidance of Richard Parker, formerly of the Medical Statistician at the Centre for Applied Medical Statistics, Department of Public Health and Primary Care, University of Cambridge, UK.

Supplementary material

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.joca.2014.03.008.

References

1. Williams FM, Spector TD. Biomarkers in osteoarthritis. Arthritis Res Ther 2008;10(1):101.
2. Lotz M, Martel-Pelletier J, Christiansen C, Brandi ML, Bruyère O, Chapurlat R, *et al.* Value of biomarkers in

osteoarthritis: current status and perspectives. Ann Rheum Dis 2013;72(11):1756–63.
3. Hunter DJ, Guermazi A, Roemer F, Zhang Y, Neogi T. Structural correlates of pain in joints with osteoarthritis. Osteoarthritis Cartilage 2013;21(9):1170–8.
4. Javaid MK, Lynch JA, Tolstykh I, Guermazi A, Roemer F, Aliabadi P, *et al*. Pre-radiographic MRI findings are associated with onset of knee symptoms: the most study. Osteoarthritis Cartilage 2010;18(3):323–8.
5. Henrotin Y. Osteoarthritis year 2011 in review: biochemical markers of osteoarthritis: an overview of research and initiatives. Osteoarthritis Cartilage 2012;20(3):215–7.
6. Turmezei TD, Poole KE. Computed tomography of subchondral bone and osteophytes in hip osteoarthritis: the shape of things to come? Front Endocrinol (Lausanne) 2011:297.
7. Bousson V, Lowitz T, Laouisset L, Engelke K, Laredo JD. CT imaging for the investigation of subchondral bone in knee osteoarthritis. Osteoporos Int 2012;23(Suppl 8):S861–5.
8. Staub D, Murphy MJ. A digitally reconstructed radiograph algorithm calculated from first principles. Med Phys 2013;40(1). 011902.
9. Arden NK, Lane NE, Parimi N, Javaid KM, Lui LY, Hochberg MC, *et al*. Defining incident radiographic hip osteoarthritis for epidemiologic studies in women. Arthritis Rheum 2009;60(4):1052–9.
10. Kellgren JH, Lawrence JS. Radiological assessment of osteoarthrosis. Ann Rheum Dis 1957;16(4):494–502.
11. Croft P, Cooper C, Wickham C, Coggon D. Defining osteoarthritis of the hip for epidemiologic studies. Am J Epidemiol 1990;132(3):514–22.
12. Reijman M, Hazes JM, Koes BW, Verhagen AP, Bierma-Zeinstra SM. Validity, reliability, and applicability of seven definitions of hip osteoarthritis used in epidemiological studies: a systematic appraisal. Ann Rheum Dis 2004;63(3):226–32.
13. Schiphof D, Boers M, Bierma-Zeinstra SM. Differences in descriptions of Kellgren and Lawrence grades of knee osteoarthritis. Ann Rheum Dis 2008;67(7):1034–6.
14. Dagenais S, Garbedian S, Wai EK. Systematic review of the prevalence of radiographic primary hip osteoarthritis. Clin Orthop Relat Res 2009;467(3):623–37.
15. Kellgren JH, Jeffrey MR, Ball J. In: The Epidemiology of Chronic Rheumatism, vol. I. Oxford: Blackwell Scientific Publications; 1963.
16. Kellgren JH, Jeffrey MR, Ball J. In: The Epidemiology of Chronic Rheumatism, vol. II. Oxford: Blackwell Scientific Publications; 1963.
17. Ornetti P, Brandt K, Hellio-Le Graverand MP, Hochberg M, Hunter DJ, Kloppenburg M, *et al*. OARSI-OMERACT definition of relevant radiological progression in hip/knee osteoarthritis. Osteoarthritis Cartilage 2009;17(7):856–63.
18. Chu Miow Lin D, Reichmann WM, Gossec L, Losina E, Conaghan PG, Maillefert JF. Validity and responsiveness of radiographic joint space width metric measurement in hip osteoarthritis: a systematic review. Osteoarthritis Cartilage 2011;19(5):543–9.
19. Treece GM, Poole KE, Gee AH. Imaging the femoral cortex: thickness, density and mass from clinical CT. Med Image Anal 2012;16(5):952–65.
20. Poole KE, Treece GM, Mayhew PM, Vaculík J, Dungl P, Horák M, *et al*. Cortical thickness mapping to identify focal osteoporosis in patients with hip fracture. PLoS ONE 2012;7(6):e38466.
21. Turmezei TD, Treece GM, Gee AH, Poole KES. Cortical thickness mapping of the proximal femur: towards a new imaging biomarker of osteoarthritis. Osteoarthritis Cartilage 2013;21:S189.
22. Ledingham J, Dawson S, Preston B, Milligan G, Doherty M. Radiographic patterns and associations of osteoarthritis of the hip. Ann Rheum Dis 1992;51(10):1111–6.
23. Solomon L, Schnitzler CM, Browett JP. Osteoarthritis of the hip: the patient behind the disease. Ann Rheum Dis 1982;41(2):118–25.
24. Li X, Benjamin Ma C, Link TM, Castillo DD, Blumenkrantz G, Lozano J, *et al*. In vivo T(1rho) and T(2) mapping of articular cartilage in osteoarthritis of the knee using 3T MRI. Osteoarthritis Cartilage 2007;15(7):789–97.
25. Watanabe A, Boesch C, Siebenrock K, Obata T, Anderson SE. T2 mapping of hip articular cartilage in healthy volunteers at 3T: a study of topographic variation. J Magn Reson Imaging 2007;26(1):165–71.
26. Nishii T, Kuroda K, Matsuoka Y, Sahara T, Yoshikawa H. Change in knee cartilage T2 in response to mechanical loading. J Magn Reson Imaging 2008;28(1):175–80.
27. Nishii T, Tanaka H, Sugano N, Sakai T, Hananouchi T, Yoshikawa H. Evaluation of cartilage matrix disorders by T2 relaxation time in patients with hip dysplasia. Osteoarthritis Cartilage 2008;16(2):227–33.
28. Choi JA, Gold GE. MR imaging of articular cartilage physiology. Magn Reson Imaging Clin N Am 2011;19(2):249–82.
29. Wong CS, Yan CH, Gong NJ, Li T, Chan Q, Chu YC. Imaging biomarker with T1ρ and T2 mappings in osteoarthritis – in vivo human articular cartilage study. Eur J Radiol 2013;82(4):647–50.
30. Lim YW, van Riet RP, Mittal R, Bain GI. Pattern of osteophyte distribution in primary osteoarthritis of the elbow. J Shoulder Elbow Surg 2008;17(6):963–6.
31. McErlain DD, Milner JS, Ivanov TG, Jencikova-Celerin L, Pollmann SI, Holdsworth DW. Subchondral cysts create increased intra-osseous stress in early knee OA: a finite element analysis using simulated lesions. Bone 2011;48(3):639–46.
32. Chiba K, Ito M, Osaki M, Uetani M, Shindo H. In vivo structural analysis of subchondral trabecular bone in osteoarthritis of the hip using multi-detector row CT. Osteoarthritis Cartilage 2011;19(2):180–5.
33. Agricola R, Waarsing JH, Arden NK, Carr AJ, Bierma-Zeinstra SM, Thomas GE, *et al*. Cam impingement of the hip – a risk factor for hip osteoarthritis. Nat Rev Rheumatol 2013;9(10):630–4.
34. Hunter DJ. Risk stratification for knee osteoarthritis progression: a narrative review. Osteoarthritis Cartilage 2009;17(11):1402–7.
35. Turmezei TD, Lomas DJ, Hopper MA, Poole KE. Severity mapping of the proximal femur: a new method for assessing hip osteoarthritis with computed tomography. Osteoarthritis Cartilage 2014;22(10):1488–98.