Reproducibility and Diagnostic Accuracy of Kellgren-Lawrence Grading for Osteoarthritis using Radiographs and Dual Energy X-ray Absorptiometry (DXA) images

Running Title: Osteoarthritis grading using DXA

Kanako Yoshida[1], Rebecca J. Barr[1], Sandro Galea-Soler[2], Richard M. Aspden[1], David M. Reid[1], and Jennifer S. Gregory*[1]

[1]Musculoskeletal Research Programme, Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK; and [2]Medical Imaging Department, Mater Dei Hospital, Msida, Malta

Abstract

Introduction: Advances in image quality from modern Dual Energy X-ray absorptiometry (DXA) scanners now allow near radiograph-like quality images at a low radiation dose. This opens potential new applications for the use of DXA scanners to study other musculoskeletal conditions, such as osteoarthritis which is often investigated by visual assessment of radiographs.
Together, osteoporosis and osteoarthritis are the two most common musculoskeletal conditions, both of which primarily affect older people. The aim of this study was to determine whether Kellgren-Lawrence grading of DXA images can be used to grade hip osteoarthritis as effectively as radiographs.

Methodology: People who had attended for recent pelvic radiographs underwent DXA images of hips (50 hips from 25 people) using a GE Healthcare iDXA scanner. Three observers assigned Kellgren-Lawrence grades to each image and grading was repeated at least one week apart. Intra-observer and inter-observer reliability for radiograph and DXA were calculated using quadratic weighted kappa (QWK). People were recalled 12 months later and the tests were repeated with both the radiograph and DXA scans taken within 2 weeks of each other.

Results: Hip DXA intra-observer reproducibility achieved a QWK range of 0.88-0.95 and inter-observer reproducibility of 0.85-0.88, similar to QWK from hip radiographs. Intra-observer reliability between subject-matched radiograph and iDXA images revealed QWK ranging between 0.80-0.88.

Conclusions: Reproducibility of hip osteoarthritis grading using DXA was comparable with that of radiographs in this study and similar to repeatability scores previously published in literature. Given the lower radiation dose and the opportunity to simultaneously investigate osteoporosis, DXA presents an attractive imaging option for osteoarthritis.

Keywords: DXA; Osteoarthritis; Osteoporosis; Bone Mineral Density; Arthritis; Kellgren-Lawrence

INTRODUCTION

Osteoporosis (OP) and osteoarthritis (OA) are the two most common musculoskeletal disorders in the developed world. Although an inverse relationship between OP and OA has been suggested (1), the diseases can coexist (2, 3) and it would be attractive to be able to use a single imaging modality to assess both in the same site (4). Modern DXA scanners have a lower radiation dose than radiographs (5-56 µSv for hip DXA (5, 6), 700 µSv for pelvic or hip (7, 8) radiographs) yet have decent image resolution, allowing assessment of vertebral fractures and aortic calcification (9).

Following the authors' observation that typical features of hip OA, osteophytes, sclerosis and joint space narrowing were clearly visible on Dual Energy Absorptiometry (DXA) images acquired for diagnosing OP by measuring Bone Mineral Density (BMD), this study investigated whether Kellgren-Lawrence grading (KLG), a standard radiographic technique for assessing osteoarthritis severity using plain films (10), can be applied reliably to DXA images.

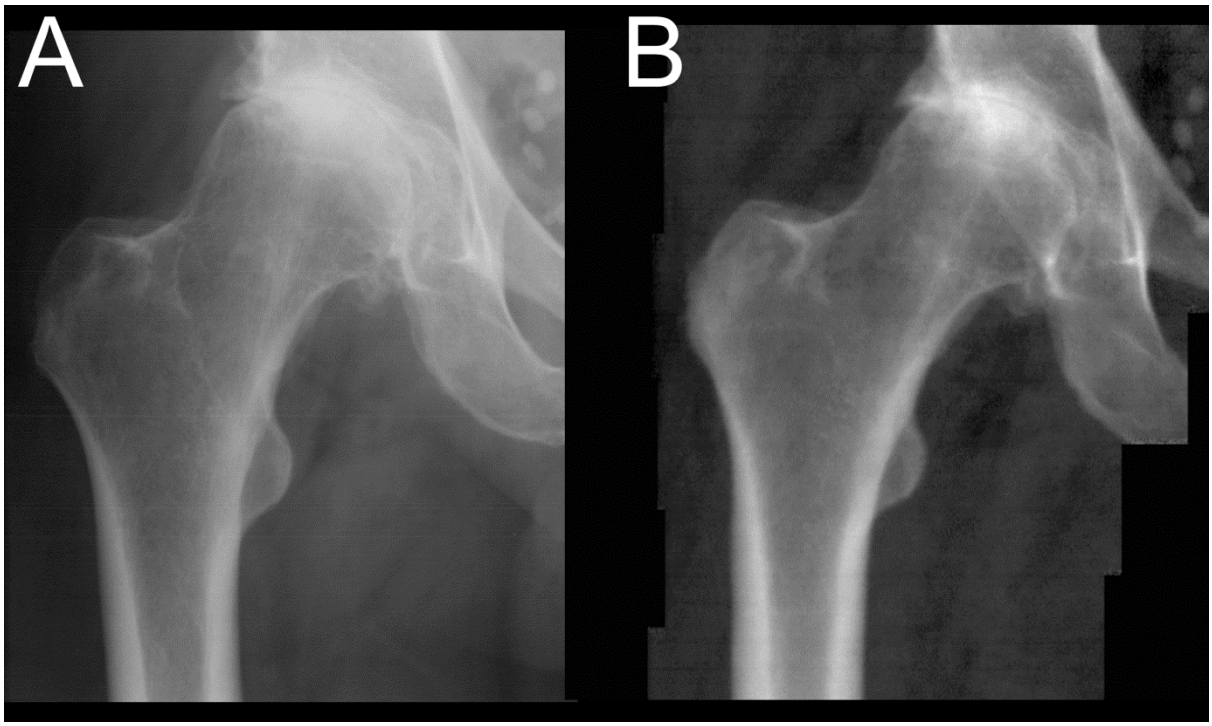


Figure 1: Comparison between radiograph and DXA images. (A) plain radiograph of the hip (B) iDXA image of the same hip. The scale and contrast of images have been adjusted for viewing purposes.

MATERIALS AND METHODS

Recruitment

Subjects for this study were identified from a larger, longitudinal study investigating osteoarthritis. Subjects for the parent study were recruited with differing degrees of hip OA identified from the local National Health Service (NHS) Radiology Information System (RIS). All patients over 30 with bilateral hip/pelvis radiographs taken within the previous year were identified via five computerised searches (April-October 2007). Based on the radiology reports (aged over 30 years, with a plain pelvic or antero-posterior radiographs of hips or knees taken on or after 1st February 2006 in any speciality except Accident and Emergency (A&E), invitation letters were sent to potential participants via their referring clinician. Radiographs were then examined for suitability. The following exclusion criteria were applied: prior surgical interventions such as total hip replacements (THR), known skeletal metastases, infective or inflammatory arthropathies, congenital/developmental dysplasia, avascular necrosis, fractures/dislocations, other bone disease (e.g. Paget's disease), or absence of a formal radiology report.

For eligible subjects who gave informed consent, DXA scans of both hips were obtained posteroanteriorly, using an iDXA scanner (GE Healthcare, Madison, WI, USA), using standard DXA positioning protocols. As part of the longitudinal study, twelve months later, they were invited for a repeat DXA scan and non-weight-bearing antero-posterior radiographs of the pelvis were also obtained. Baseline images were used for initial comparison of KLG. Results were later confirmed using the 12-month images where DXA and radiographic images were taken within 1 week.

A subset of baseline radiographs (50 hips, 25 subjects) with subject-matched DXA scans, encompassing the full KLG range (0-4) was selected for this reproducibility study by JSG who was not involved in grading. Radiographs and DXA images were graded independently and in random order by 3 observers (KY, SG-S, DMR) from rheumatology or radiology backgrounds at consultant and trainee level. The images were graded again, randomly and independently, at least one week later without knowledge of the previous grades by the same observers. To enable off-site scoring radiographic images from both visits were digitized for DMR, whereas only the second set was digitised for the other observers. Radiographs were digitized using a Howtek MultiRAD 850 (Howtek, Hudson, New Hampshire) at 146 dpi and 8-bit depth. Observers could identify left and right hip images from the same patients; no other subject identifiable information was available.

DXA and radiographic images from the second visit were graded by KY (twice) and SG-S (once). Six subjects from the original 50 either had a THR or withdrew before this visit. These were replaced with subjects of similar age and baseline KLG to ensure QWK statistics were directly comparable.

Table 1: Distribution of baseline radiographic KLG in the study.

Grade recorded	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4	TOTAL
Pelvic X-ray	5	16	13	7	9	50

Note: Modes of all grades for each image by the three observers were used, unless more than one mode was possible, where the median was rounded to the nearest integer.

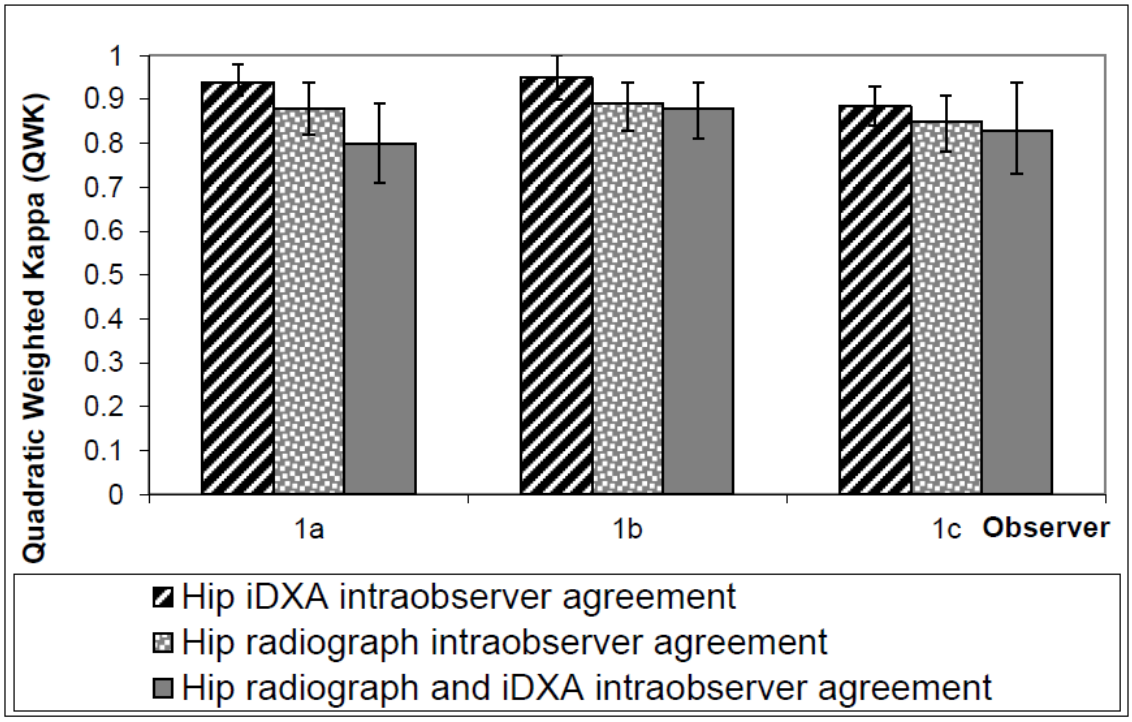


Figure 2: Intra-observer reproducibility for iDXA and radiograph of the Hip. Error bar represents 95% confidence interval (based on standard error, $H_0 \neq 0$)

Osteoarthritis grading

Images were graded according to the Kellgren and Lawrence system using a reprint of the *Atlas of standard radiographs of arthritis* (11) and the *Atlas of individual radiographic features, revised* (12) (for KLG 0, since a "normal" image was not included in the original atlas). Observers were permitted to alter contrast and magnification of digital images using ImageJ.

Statistics

Reliability was calculated using Quadratic Weighted-Kappa (QWK) using MedCalc (v9.4.1, MedCalc, Mariakerke, Belgium) and WINPEPI (v9.3, PAIRSetc) and Intraclass Correlation Coefficients (ICC) using SPSS (v17, SPSS Inc, Chicago, USA) with two way random and absolute agreement**.**

A Kappa score of "1" indicates perfect agreement, "0" chance and "-1" perfect disagreement. Kappa is suitable for dichotomous or unordered categorical variables. When categories have a ranking or order, such as KLG, weighted Kappa is more appropriate (13). For comparison with previous studies, we also calculated the ICC which, while most appropriately used for continuous variables, is equivalent to QWK (14).

RESULTS

The study comprised 12 men and 13 women, average age 65.9 (±9.3) years. The average interval between the recruitment radiograph and baseline DXA was 225 (±104) days. All 12-month DXAs and radiographs were taken within 1 week. Figure 1 shows a typical osteoarthritic DXA and radiograph. Table 1 shows the baseline radiographic KLG distribution. There were no adverse events from performing the radiograph or DXA scan.

Intra-observer and inter-observer reproducibility

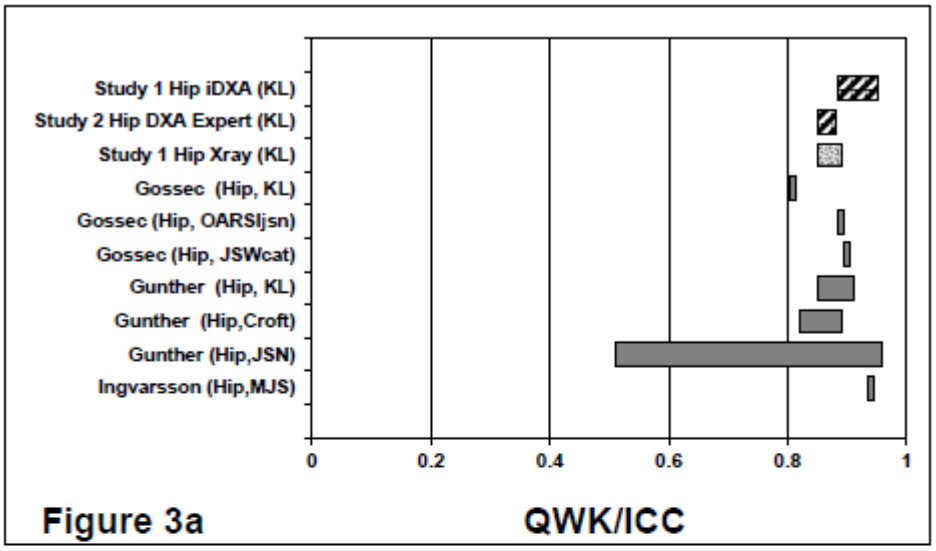
Good levels of intra-observer (Figure 2) and inter-observer (Table 2) reproducibility were achieved. All observers had similar intra-observer QWK values of 0.88-0.95 and 0.85-0.88 for DXA and radiographs respectively, and corresponding absolute agreements of 68-88% and 62-66%. Inter-observer agreement for DXA images was also similar (QWK 0.85-0.88, Table 2).

Calculation of ICC for all values confirmed QWK approximated to ICC (14) with a difference of no more than 0.01 for both two-way random and two-way mixed effect models.

Table 2: Inter-observer reproducibility scores for each pair of observers

		Absolute agreement %		QWK (SE)		ICC (CI)
		Obs 2	Obs 3	Obs 2	Obs 3	
Hip iDXA	Obs 1	60%	58%	0.86 (0.037)	0.88 (0.027)	0.86 (0.79-0.92)
	Obs 2		56%		0.85 (0.035)	
Hip radiograph	Obs 1	54%	34%	0.81 (0.048)	0.67 (0.067)	0.75 (0.63-0.84)
	Obs 2		52%		0.73 (0.063)	

Abbr: CI: 95% confidence interval; ICC: Intraclass correlation coefficient; Obs, observer, QWK: quadratic weighted kappa. SE: Standard error ($H_0 \neq 0$).



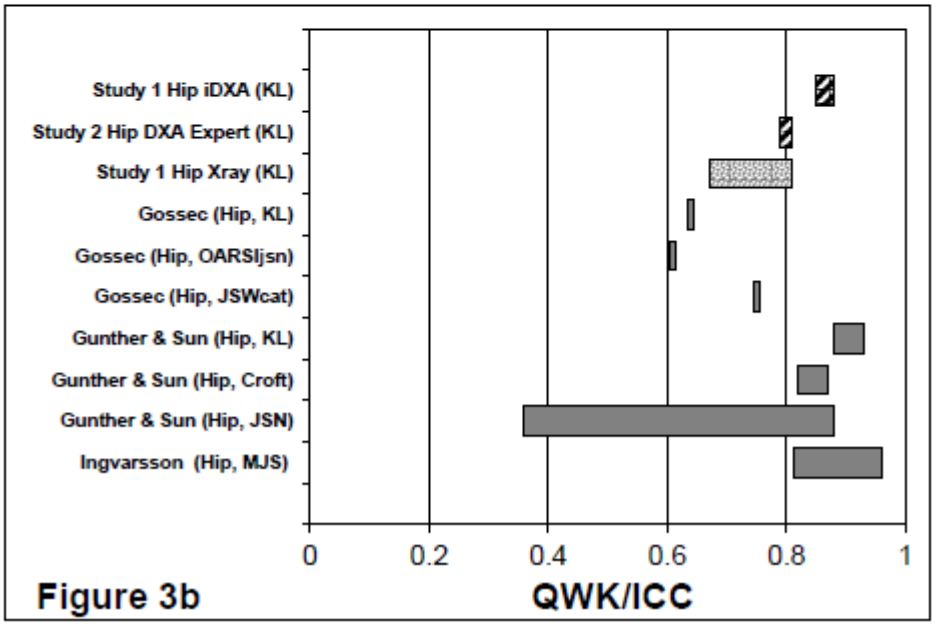


Figure 3: Comparison of study results to literature. Comparison of results from the current study measuring intra-observer (**3a**) and inter-observer reliability (**3b**) from KLG of baseline DXA images (striped black) and radiographs (dotted grey) with published data from other studies (plain grey) that measured the reliability of OA grading of the hip using ICC or QWK (18, 19, 27). For Gossec et al. (18) Ingvarsson et al. (28), as only one ICC value was obtained, the width of the bar was widened to 0.01 for visual ease. For joint space narrowing (JSN) in Günther and Sun (27), the ICC range includes both superior and medial JSN. Joint space width (JSW) was categorised to measure JSW (JSWcat) (18).

DXA vs. radiograph

Intra-observer agreement between radiographs and DXAs achieved QWK values of 0.80-0.88 (Figure 2) with no mode or median grades differing by more than 1 KLG.

12 month time-point

Intra-observer reliability for KY was 0.99 (SE 0.005; 95% CI 0.99-1) for DXA and 0.88 (SE: 0.03; 95% CI 0.82-0.94) for radiographs; inter-observer reliability between KY and SGS was 0.87 (SE 0.024; 95% CI 0.83-0.92) for DXA and 0.85 (SE 0.037; 95% CI 0.78-0.93) for radiographs. The intra-

observer QWK values of 0.89 and 0.95 between DXA and radiograph images were similar to, but slightly higher than baseline grades for these observers.

To put our results in context we searched the literature for studies reporting radiographic OA reliability. Figure 3 shows ICC and QWK values from published studies looking at KLG, Croft, Osteoarthritis Research Society international (OARSI) grading and joint space width compared to the current study.

DISCUSSION

These results demonstrate that KLG can be applied to DXA images of the hip from iDXA scanners as reproducibly as standard radiographs and the same grade was assigned to the majority of subjects, regardless of the image source. All QWK scores lay in, or above the 0.61–0.80 range, referred to as 'good' (15) or 'substantial' (16) agreement.

A simple classification of 'good' agreement using a cut-off value can be considered to be somewhat arbitrary (16, 17) and does not fully evaluate the strength of DXA imaging for OA. However as Figure 3 demonstrates, DXA scoring was at least as good as radiographs and figures in the literature. Unfortunately, it is difficult to compare with published results in more depth as there is often a lack of detail of OA severity or prevalence, and of the statistical tests used, for example specifying the version of kappa (weighted/unweighted, Fleiss-Cohen/Cichetti-Allison weights), or ICC (McGraw, Shrout-Fleiss's 6 types), or the statistical package.

This study suffers from some limitations. Although comparable to many (18, 19), the number of subjects is still relatively small, and only basic randomisation (the order of images as presented to graders) was achieved. In addition, the study did not include a consensus session to discuss images where there was disagreement. This should affect DXA and radiographs equally and not add bias but may have reduced inter-observer repeatability.

Intra-observer scores comparing radiographs and DXA were "good", though unsurprisingly slightly lower than for each modality alone, probably because of differing contrast and resolution. Although slightly higher, 12-month repeatability was similar to baseline, indicating that the gap between radiograph and DXA acquisition at baseline (225 days vs. < 7 days), caused by using historical radiographs taken as part of the subject's normal healthcare for recruitment, had little impact on KL repeatability.

DXA has some advantages compared with radiographs, including the low radiation dose, measurement of BMD and the use of positioning devices and strict protocols as standard practice (recent testing of our radiographers using 60 volunteers gave a precision error of 0.72% and least significant change 2.0%, less than half those recommended by the International Society for Clinical Densitometry (ISCD) (20)). Disadvantages of DXA include lower resolution and the inability to take weight-bearing images, so joint space measurements cannot currently be as precise as on radiographs.

Osteoarthritis is a complex disease where there is limited concordance between symptoms and radiographic features and links between them are not fully understood (21). Whilst individuals may often be diagnosed on symptoms alone (22), structural changes fundamentally underpin disease progression and are also critical for complete understanding of osteoarthritis, particularly for evaluation of therapeutic agents.

CONCLUSION

This study has shown hip KLG on iDXA images is at least as reproducible as on radiographs. Our results were comparable with published literature and it may be that the use of positioning devices in DXA minimises variability. The ability to use DXA images to assess radiographical OA will create further clinical and research opportunities, although testing would be recommended for each manufacturer and scanner model. The relative accessibility and lower radiation dose of DXA makes the technology an appealing modality. Furthermore, as the elderly population at risk of both osteoporosis and osteoarthritis expands and therapeutic agents effective in osteoporosis show promise for use in osteoarthritis (23-26), the potential for a one-stop scan to assess both diseases makes DXA an attractive modality to consider for use in standard clinical practice.

ACKNOWLEDGEMENTS

The authors would like to thank all the study volunteers, radiographers Lana Gibson and Jennifer Scott as well as Carol McKerron for administrative support. The acquisition of scans was supported in part by an award (Ref: WHMSB_AU_068_071) from the Translational Medicine Research Initiative - a consortium made up of the Universities of Aberdeen, Dundee, Edinburgh and Glasgow, the four associated NHS Health Boards (Grampian, Tayside, Lothian and Greater Glasgow & Clyde), Scottish Enterprise and Pfizer. Kanako Yoshida is funded by the Grampian Osteoporosis Trust (GOT). Jennifer Gregory is supported by the Medical Research Council [G0901242].

REFERENCES

1. Foss MV, Byers PD. 1972 Bone density, osteoarthrosis of the hip, and fracture of the upper end of the femur. Ann Rheum Dis. 31(4):259-264.

2. Healey JH, Vigorita VJ, Lane JM. 1985 The coexistence and characteristics of osteoarthritis and osteoporosis. J Bone Joint Surg Am. 67(4):586-592.

3. Mäkinen TJ, Alm JJ, Laine H, Svedström E, Aro HT. 2007 The incidence of osteopenia and osteoporosis in women with hip osteoarthritis scheduled for cementless total joint replacement. Bone. 40(4):1041-1047.

4. Johnell O, Kanis JA. 2006 An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int. 17(12):1726-1733.

5. Radiation protection—doses and Legislation. In: 2007 Fundamentals of Bone Densitometry. CD Edition. Anonymous National Osteoporosis Society, Bath, UK.

6. Steel SA, Baker AJ, Saunderson JR. 1998 An assessment of the radiation dose to patients and staff from a lunar expert-XL fan beam densitometer. Physiol Meas. 19(1):17-26.

7. Wall BF, Hart D. 1997 Revised radiation doses for typical X-ray examinations: Report on a recent review of doses to patients from medical X-ray examinations in the UK by NRPB. Br J Radiol. 70(MAY):437-439.

8. Mettler Jr. FA, Huda W, Yoshizumi TT, Mahesh M. 2008 Effective doses in radiology and diagnostic nuclear medicine: A catalog. Radiology. 248(1):254-263.

9. Rea JA, Li J, Blake GM, Steiger P, Genant HK, Fogelman I. 2000 Visual assessment of vertebral deformity by X-ray absorptiometry: A highly predictive method to exclude vertebral deformity. Osteoporos Int. 11(8):660-668.

10. Kellgren JH, Lawrence JS. 1957 Radiological assessment of osteo-arthrosis. Ann Rheum Dis. 16(4):494-502.

11. The atlas of standard radiographs of arthritis. 2005 Rheumatology (Oxford). 44 Suppl 4:iv46-iv72.

12. Altman RD, Gold GE. 2007 Atlas of individual radiographic features in osteoarthritis, revised. Osteoarthritis Cartilage. 15(SUPPL. 1):1-56.

13. Sim J, Wright CC. 2005 The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Phys Ther. 85(3):257-268.

14. Fleiss JL, Cohen J. 1973 The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas. 33:613-619.

15. Altman DG. 1991 Some Common Problems in Medical Research. In: Practical Statistics for Medical Research. Anonymous Chapman and Hall, London: 396-439.

16. Landis JR, Koch GG. 1977 The measurement of observer agreement for categorical data. Biometrics. 33(1):159-174.

17. Ludbrook J. 2002 Statistical techniques for comparing measurers and methods of measurement: A critical review. Clin Exp Pharmacol Physiol. 29(7):527-536.

18. Gossec L, Jordan JM, Lam MA, et al. 2009 Comparative evaluation of three semi-quantitative radiographic grading techniques for hip osteoarthritis in terms of validity and reproducibility in 1404 radiographs: Report of the OARSI-OMERACT task force. Osteoarthritis Cartilage. 17(2):182-187.

19. Reijman M, Hazes JMW, Koes BW, Verhagen AP, Bierma-Zeinstra SMA. 2004 Validity, reliability, and applicability of seven definitions of hip osteoarthritis used in epidemiological studies: A systematic appraisal. Ann Rheum Dis. 63(3):226-232.

20. Baim S, Wilson CR, Lewiecki EM, Luckey MM, Downs Jr. RW, Lentle BC. 2005 Precision assessment and radiation safety for dual-energy X-ray absorptiometry: Position paper of the international society for clinical densitometry. Journal of clinical densitometry : the official journal of the International Society for Clinical Densitometry. 8(4):371-378.

21. Reijman M, Hazes JMW, Koes BW, Verhagen AP, Bierma-Zeinstra SMA. 2004 Validity, reliability, and applicability of seven definitions of hip osteoarthritis used in epidemiological studies: A systematic appraisal.. Ann Rheum Dis. 63(3):226-232.

22. National institute for health and care excellence (osteoarthritis: Care and management in adults). London: National Institute for Health and Care Excellence; 2014. Report No.: [CG177].

23. Bruyere O, Delferriere D, Roux C, et al. 2008 Effects of strontium ranelate on spinal osteoarthritis progression. Ann Rheum Dis. 67(3):335-339.

24. Laslett LL, Doré DA, Quinn SJ, et al. 2012 Zoledronic acid reduces knee pain and bone marrow lesions over 1 year: A randomised controlled trial. Ann Rheum Dis. 71:1322-1328.

25. Reginster JY, Chapurlat R, Christiansen C, et al. 2012 Structure modifying effects of strontium ranelate in knee osteoarthritis. Osteoporos Int. 23 (Suppl 2):S58: OC3.

26. Esenyel M, Içağasioğlu A, Esenyel CZ. 2012 Effects of calcitonin on knee osteoarthritis and quality of life. Rheumatology International. 33:1-5.

27. Günther KP, Yi S. 1999 Reliability of radiographic assessment in hip and knee osteoarthritis. Osteoarthritis Cartilage. 7(2):239-246.

28. Ingvarsson T, Hagglund G, Lindberg H, Lohmander LS. 2000 Assessment of primary hip osteoarthritis: Comparison of radiographic methods using colon radiographs. Ann Rheum Dis. 59(8):650-653.