Osteoarthritis and Cartilage 22 (2014) 969–974

Osteoarthritis and Cartilage

Inter-observer reliability for radiographic assessment of early osteoarthritis features: the CHECK (cohort hip and cohort knee) study

J. Damen † *, D. Schiphof †, S. Ten Wolde ||, H.A. Cats ¶, S.M.A. Bierma-Zeinstra † ‡, E.H.G. Oei §

† Department of General Practice, Erasmus MC, Rotterdam, The Netherlands
‡ Department of Orthopaedics, Erasmus MC, Rotterdam, The Netherlands
§ Department of Radiology, Erasmus MC, Rotterdam, The Netherlands
|| Department of Rheumatology, Kennemer Gasthuis Haarlem, Haarlem, The Netherlands
¶ Department of Rheumatology, St. Maartenskliniek, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 9 October 2013
Accepted 7 May 2014

Keywords:
Osteoarthritis
Radiography
Inter-observer reliability

SUMMARY

Objective: To calculate inter-observer reliability between four different trained readers and an experienced reader on early radiographic osteoarthritis (OA) features in our early OA cohort hip and cohort knee (CHECK) cohort.

Methods: Four readers were trained by a radiologist and experienced reader to score radiographic OA features. After this training they scored the CHECK cohort. Of the 1002 participants, 38 were scored by all readers. Five different angle radiographs (three for the knee, two for the hip) at three different time points were scored and compared. Inter-observer reliability was evaluated between each of the four trained readers and the experienced reader. Separate radiographic OA features and of overall Kellgren & Lawrence (K&L) scores. In addition, reliability of progression of radiographic was determined in K&L scores and joint space narrowing (JSN).

Results: For hip and knee there was substantial inter-observer reliability on overall K&L scores. In the knee, JSN was scored with fair to moderate reliability, osteophytes with moderate to nearly perfect reliability, and other features with fair to substantial reliability. In the hip, reliability ranged from substantial to nearly perfect. Moderate inter-observer reliability was found for progression of OA in both knee and hip, with slightly better reliability for progression based on K&L scores than on separate features.

Conclusion: Good inter-observer reliability can be achieved between trained readers and an experienced reader. Although JSN in the knee is scored with lower inter-observer reliability than osteophytes, this does not seem to influence overall K&L scoring. In the hip all features showed good reliability.

Introduction

Radiographic assessment of osteoarthritis (OA) is very important, as findings on radiography are used for both diagnosis and staging of OA and as a tool for measuring progression of disease. Despite limitations of radiography and the emergent role of magnetic resonance imaging (MRI) for OA, radiography is still the most widely applied imaging tool for OA in clinical and research settings. This is mainly because of the wide availability of the technique, low costs, and relatively easy interpretation of the images[1].

Many studies have indicated the discordance between radiological findings and symptoms pertaining to OA[2,3]. Several approaches have been suggested to improve the diagnostic and prognostic value of radiographic findings, including redefinition of existing grading systems, implementing new quantitative grading methods with actual measurements, and performing additional angled radiographic views to find relationships between pain and radiographic joint damage[4–6].

A common problem in all these approaches is the difficulty of standardized classification of OA. While standardized acquisition of radiography is commonly applied, the method of grading of OA features, OA definition varies by study and training of readers. These factors make it even more difficult to relate OA signs to clinical outcomes, because data from different studies cannot easily be combined.

* Address correspondence and reprint requests to: J. Damen, Erasmus MC, Dept. of General Practice, Room GK-1044, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands. Tel: 31-10-7043815; Fax: 31-10-7044766.
E-mail address: j.damen@erasmusmc.nl (J. Damen).

http://dx.doi.org/10.1016/j.joca.2014.05.007
1063-4584/

Assessment and improvement of inter-observer reliability might contribute to solving this problem as it leads to more uniformity in radiographic scoring both between and within studies. However, compared to the large number of studies performed using radiological outcomes, not all studies have reported on their inter reliability of OA grading. We found only one review article on inter-observer reliability for radiographic grading of hip and knee OA, which indicated that only few data exist on this issue, especially for separate radiological OA features; therefore, the authors advised that in larger cohort studies inter-observer reliability should be reported[7].

The purpose of the present study was to assess various aspects of inter-observer reliability for radiographic hip and knee OA assessment within the longitudinal cohort hip and cohort knee (CHECK) study. First, we aimed at evaluating inter-observer reliability between four trained readers and one experienced reader for grading of radiographic OA, including analyses per separate OA feature. Second, we aimed to determine the inter-observer reliability for assessing progression of OA, which is relevant for longitudinal studies on OA.

Methods

CHECK is an ongoing prospective multicentre cohort study of 1002 individuals with early symptomatic OA of the knee or hip. Details of the protocol have been published earlier, and a summary is presented below[8].

Study population

Participants that potentially fulfilled the inclusion criteria were invited to join the study when they visited their general practitioner (GP). In addition, participants were recruited through advertisements, articles in local newspapers, and via the website of the Dutch Arthritis Association. Individuals were eligible to participate if they had pain and/or stiffness of the knee and/or hip, were aged between 45 and 65 years, and had not yet consulted their physician for these symptoms, or the first consultation was within 6 months before entry.

Exclusion criteria were: any other pathological condition that could explain the existing symptoms (e.g., other rheumatic disease, isolated tendinitis/bursitis, previous hip or knee joint replacement, congenital dysplasia of the hip, osteochondritis dissecans, intra-articular fractures, septic arthritis, Perthes' disease, ligament or meniscus damage, Plica syndrome or Bakers' cysts), or a co-morbidity that precluded physical evaluation and/or follow-up for at least 10 years, malignancy in the last 5 years, and inability to understand the Dutch language.

Physicians at the participating centres checked whether referred participants and participants reacting to advertisements, as well as participants from their own outpatient clinic, fulfilled the inclusion criteria.

All patients underwent radiographic assessment of both hips and knees, standardised physical examination, and filled out an extensive questionnaire at baseline, and at 2 and 5 years follow-up, including the Western Ontario McMaster questionnaire (WOMAC) pain subscale[8,9].

Radiography

All radiographs in the CHECK study were obtained according to a standardized protocol: Semi-flexed (7–10°) weight-bearing posteroanterior (PA) radiographs of the tibiofemoral joints were made, followed by standing mediolateral views in 30° flexion for assessment of the tibiofemoral and patellofemoral joints. Skyline (infero-superior) non-weight bearing views of the patellofemoral joints were also acquired (with 30° flexion of the femorotibial joints). For the hip, weight-bearing anteroposterior (AP) radiographs of the pelvis were made, as well as weight-bearing faux profile radiographs of both hips taken according to Lequesne and Loredo[10]. The faux profile view provides a lateral projection of the femoral head and neck, and an oblique view of the acetabulum tangential to its superoanteromedial edge[10].

Radiographic OA scoring and training

When readers scored approximately 75 participants each, a subset of 38 participants from different centres was scored by five readers for the purpose of this study on inter-observer reliability. Radiographic OA scoring of this subset was performed independently by five trained observers (four research assistants and one experienced GP reader). For these random selected participants no selection criteria were used other than having no missing data. Of each participant, we scored right and left radiographs at three different points in time. Radiographs at baseline, and at 2 and 5-years follow-up were all scored at the same time and the readers were aware of their sequence in time.

Before scoring, the four research assistants were extensively trained by an experienced musculoskeletal radiologist and the experienced GP reader in four separate training sessions of 2.5 h each, using training radiographs from the CHECK cohort. At the end of this training program the assistants' performance was assessed by scoring a set of radiographs of 12 participants with differing OA severity in the presence of the GP reader. When the assistants reached good reliability compared to the GP reader they were allowed to score radiographs in the CHECK cohort.

The PA radiographs of the knee were scored for individual OA features according to Altman *et al.*[11] For grading of OA, the Kellgren & Lawrence (K&L) definition was used for the patellofemoral joint, as determined on the PA radiograph[12]. Medial and lateral joint space narrowing (JSN), femoral medial and lateral osteophytes, and tibial medial and lateral osteophytes were scored on a 0–3 scale (0 = normal; 1 = mild; 2 = moderate; and 3 = severe). Tibial bone attrition, tibial sclerosis and femoral sclerosis were scored both medial and lateral on a 0–1 scale (0 = absent; 1 = present). Spiking of the tibial spines was scored according to the atlas of Burnett *et al.* on a 0–1 scale[13]. The mediolateral and skyline radiographs of the knee were also scored for patellofemoral JSN and osteophytes on a 0–3 scale, as well as patellofemoral sclerosis (only on the skyline view) according to Burnett *et al.*[13]. For these random selected participants no selection criteria were used other than having no missing data.

The AP radiographs of the hip were scored for individual radiographic OA features according to Altman *et al.*[11] For grading of OA, the K&L definition was used and assessed on the AP radiograph[12]. Superior and medial hip JSN, superior and inferior acetabular osteophytes, and superior and inferior femoral osteophytes were scored on a 0–3 scale. Inferior acetabular osteophytes, femoral subchondral sclerosis, acetabular subchondral cysts, flattening of the femoral head and buttressing were scored on a 0–1 scale[13]. On the faux profile radiographs superior and posterior JSN were scored on a 0–3 scale.

For measuring reliability on progression we defined progression as an increase of at least one K&L grade. For JSN we defined progression as an increase of at least one point on the 0–3 scale in one of the sites. When calculating inter-observer reliability for JSN and osteophytes we used the maximum JSN and largest osteophyte scored at one of the subsites where JSN or osteophytes were scored. Progression was determined during the time intervals: baseline to