**Appendix: Reliability assessment of hand OA radiographs read singly and in series**

In clinical trials for osteoarthritis (OA) research, it is generally recommended that all radiographic films from the same person be evaluated in series (paired readings) at the same time after the final films have been obtained (1, 2). Many studies, whether observational or trials, utilize paired readings of hand radiographs, although it is not clear that such readings are necessary (3-5). In fact, Kallman et al reported high reliability of Kellgren-Lawrence grade (K-L grade) for cross-sectional, unpaired scoring of hand radiographs obtained in a longitudinal study (6). This method was also endorsed, for epidemiologic studies in particular, in a report from the second international workshop on hand OA, with the thought that random reads would reduce time-sequence bias which could result in overestimation of progression (1).

In the setting of a large longitudinal study such as the JoCo OA Project, where multiple films are obtained over many years, a requirement for reading all films simultaneously for all 30 joints in both hands becomes extremely burdensome and resource intensive, especially considering the fact that JoCo OA data collection is ongoing, and it is impractical to re-read all radiographs in series with each additional data cycle. It would be far more efficient to read each film only once, if such readings were comparably reliable. Therefore, to determine whether use of cross-sectional readings (films for each person read singly for each data collection cycle) was valid for determining radiographic OA status among participants in our study, we conducted a reliability study to compare radiographic findings from cross-sectional versus paired hand radiograph readings in a subset of JoCo OA Project participants.

**Methods**

*Definition of hand OA*

All hand radiographs were read by a single experienced reader (JBR). The films were read separately at the time when they were obtained (cross-sectional) and, for the purposes of this analysis, together in pairs in known chronologic order (paired) in a subset of radiographs (described below). Cross-sectional readings were conducted in each data collection cycle for Time 1 (1999–2004) and Time 2 (2005–2010) after the X-ray films were obtained, and paired readings for the reliability assessment were conducted in 2015. Posteroanterior radiographs of each hand were read for K-L grade (0-4) at each of the 15 joints (4 distal interphalangeal [DIP], 4 proximal interphalangeal [PIP], 4 metacarpophalangeal [MCP], 1 carpometacarpal [CMC], 1 thumb IP and 1 thumb MCP). Radiographic OA (rOA) at the joint level was defined if the joint had a K-L grade ≥ 2. We defined the presence of hand rOA at the person level as a K-L grade ≥ 2 involving at least 3 hand joints (DIP, PIP, or CMC, with the thumb IP considered a PIP), and at least one of them being a DIP joint, in at least one hand (same as in the manuscript).

*Selection of radiographs for paired readings*

We used a stratified sampling method to ensure that the subset of paired radiographs examined in this reliability study represented the range of K-L grades. First, we created a joint score as the sum of K-L grades across selected joints on the right hand at baseline (CMC, DIP 2-3, PIP 2-3, MCP 2-3; total score 0-28). Then, we stratified scores as low (0-1), medium (2-4), high (5-9), and very high (10+), and randomly selected 15 hands from each stratum yielding a sample size of 60 right hands. Next, the radiologist (JBR) conducted 60 paired hand radiograph readings (for all joints).

*Statistical analysis*

We used the Kappa coefficient (κ) to compare the K-L grade for the cross-sectional and corresponding paired readings; agreement was examined between each of the joints in each of the two data collection cycles (Time 1 and Time 2). The κ coefficient is a statistic which can be used to measure intra-rater agreement for categorical variables. It is used as a more robust measure than simple percent agreement because κ accounts for agreement occurring by chance. To interpret κ, a commonly cited scale is: <0: less than chance agreement; 0.01-0.20: slight agreement; 0.21-0.40: fair agreement; 0.41-0.60: moderate agreement; 0.61-0.80: substantial agreement; 0.81-0.99: almost perfect agreement (7). When comparing agreement in K-L grades for each of the 15 hand joints, we used the weighted κ which gives partial weight to results differing by one or more K-L grade. We used the unweighted κ when comparing participant’s hand rOA status between the two data collection cycles, as this variables is measured as present or absent.

**Results**

The 60 individuals contributing hand films to this analysis were an average age of 70.3 years, , 33.3% men, 26.7% African American, with an average body mass index (BMI) of 32.5. Appendix Table 1 details the weighted κ with corresponding 95% confidence intervals for comparing the cross-sectional versus paired reads for each of the 15 joints per hand, separately for baseline and follow-up radiographs. These demonstrate substantial to almost perfect agreement for all joints.

Appendix Table 2 displays the cross-tabulations comparing hand rOA (defined at the person level) between the cross-sectional and paired reads, separately for Time 1 and Time 2. The associated κ coefficients were 0.72 (indicating substantial agreement) and 0.86 (almost perfect agreement), respectively for Time 1 and Time 2.

**Conclusions**

These results indicate the high degree of reliability for K-L grade of each hand joint in this sample at both data collection cycles when using cross-sectional readings as compared to paired readings. The reliability of the overall definition of hand rOA, which involves these joints, is correspondingly high as well. Our results show that moderate to substantial agreement is possible, using a single experienced reader, without the necessity of pairing films over time or reading multiple films per participant in chronological order.

**References**

1. Buckland-Wright JC, Verbruggen G, Haraoui PB. Imaging. Radiological assessment of hand OA. Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society 2000;8 Suppl A:S55-6.

2. Hunter DJ, Arden N, Cicuttini F, et al. OARSI Clinical Trials Recommendations: Hand imaging in clinical trials in osteoarthritis. Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society 2015;23(5):732-46.

3. Haugen IK, Englund M, Aliabadi P, et al. Prevalence, incidence and progression of hand osteoarthritis in the general population: the Framingham Osteoarthritis Study. Ann Rheum Dis 2011;70(9):1581-6.

4. Maheu E, Cadet C, Gueneugues S, et al. Reproducibility and sensitivity to change of four scoring methods for the radiological assessment of osteoarthritis of the hand. Annals of the rheumatic diseases 2007;66(4):464-9.

5. Bijsterbosch J, Haugen IK, Malines C, et al. Reliability, sensitivity to change and feasibility of three radiographic scoring methods for hand osteoarthritis. Annals of the rheumatic diseases 2011;70(8):1465-7.

6. Kallman DA, Wigley FM, Scott WW, Jr., et al. New radiographic grading scales for osteoarthritis of the hand. Reliability for determining prevalence and progression. Arthritis and rheumatism 1989;32(12):1584-91.

7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-74.

Appendix Table 1. Weighted κ coefficients and 95% confidence intervals (CI) for the K-L grade in each hand joint comparing cross-sectional versus paired reads, separately for Time 1 and Time 2 (n=60)

|  |  |  |
| --- | --- | --- |
|  | Time 1 | Time 2 |
|  | Weighted k (95% CI) | Weighted k (95% CI) |
| LEFT |  |  |
| 2 DIP  | 0.78 (0.67-0.89) | 0.74 (0.64-0.84) |
| 3 DIP | 0.71 (0.58-0.85) | 0.78 (0.67-0.89) |
| 4 DIP | 0.71 (0.60-0.82) | 0.78 (0.67-0.88) |
| 5 DIP |  0.81 (0.71-0.92)\* |  0.82 (0.74-0.91)\* |
| 1 IP | 0.80 (0.69-0.92) | 0.75 (0.64-0.85) |
| 2 PIP | 0.74 (0.61-0.88) | 0.79 (0.67-0.90) |
| 3 PIP | 0.69 (0.54-0.83) | 0.73 (0.62-0.84) |
| 4 PIP | 0.71 (0.56-0.85) |  0.82 (0.72-0.92)\* |
| 5 PIP | 0.72 (0.59-0.85) | 0.77 (0.65-0.90) |
| 1 MCP |  0.49 (0.24-0.74) † | 0.67 (0.55-0.80) |
| 2 MCP | 0.67 (0.49-0.84) | 0.63 (0.46-0.80) |
| 3 MCP | 0.73 (0.60-0.87) | 0.65 (0.48-0.82) |
| 4 MCP | 0.76 (0.58-0.94) |  0.89 (0.77-1.00)\* |
| 5 MCP | 0.72 (0.45-1.00) | 0.78 (0.59-0.97) |
| 1 CMC | 0.68 (0.55-0.81) |  0.57 (0.44-0.71) † |
| RIGHT |  |  |
| 2 DIP  | 0.79 (0.68-0.89) | 0.74 (0.65-0.84) |
| 3 DIP | 0.77 (0.65-0.90) | 0.77 (0.67-0.88) |
| 4 DIP | 0.68 (0.55-0.80) | 0.77 (0.66-0.87) |
| 5 DIP | 0.79 (0.69-0.89) | 0.76 (0.67-0.85) |
| 1 IP | 0.78 (0.67-0.89) | 0.80 (0.70-0.90) |
| 2 PIP | 0.77 (0.65-0.89) | 0.77 (0.65-0.89) |
| 3 PIP |  0.58 (0.44-0.72) † |  0.84 (0.74-0.93)\* |
| 4 PIP |  0.57 (0.43-0.71) † | 0.73 (0.60-0.85) |
| 5 PIP | 0.75 (0.62-0.88) | 0.74 (0.62-0.86) |
| 1 MCP | 0.64 (0.45-0.82) | 0.61 (0.47-0.75) |
| 2 MCP | 0.69 (0.56-0.83) | 0.77 (0.64-0.90) |
| 3 MCP | 0.61 (0.41-0.80) | 0.61 (0.45-0.77) |
| 4 MCP | 0.62 (0.36-0.87) | 0.61 (0.41-0.81) |
| 5 MCP |  0.57 (0.36-0.78) † | 0.70 (0.55-0.86) |
| 1 CMC | 0.75 (0.62-0.87) | 0.80 (0.70-0.90) |
| All k indicated substantial (0.61-0.8) unless otherwise noted as moderate (0.41-0.6, †) or almost perfect (0.81 or higher, \*) |

Appendix Table 2. Cross-tabulation and κ (95% confidence interval) comparing cross-sectional and paired reads for Hand rOA defined at the person level, separately for Time 1 and Time 2 (n=60).

|  |  |
| --- | --- |
|  | **Hand rOA by paired read** |
| **Hand rOA by cross-sectional read** | Absent | Present | Total |
| **Time1** |  |  |  |
| Absent  | 39 | 4 | 43 |
| Present | 3 | 14 | 17 |
| Total | 42 | 18 | 60 |
| κ: 0.72 (0.52-0.91)|  |
| **Time2** |  |  |  |
| Absent  | 32 | 1 | 33 |
| Present | 3 | 24 | 27 |
| Total | 35 | 25 | 60 |
| κ: 0.86 (0.74-0.99) |