BMJ Open

# Clinicians' heuristic assessments of radiographs compared with Kellgren-Lawrence and Ahlbäck ordinal grading: an exploratory study of knee radiographs using paired comparisons

Mads Møller Pedersen,[1] Kristian Breds Geoffroy Mongelard,[2] Anne Mørup-Petersen,[3] Karl Bang Christensen,[1] Anders Odgaard[4,5]

**To cite:** Pedersen MM, Geoffroy Mongelard KB, Mørup-Petersen A, *et al*. Clinicians' heuristic assessments of radiographs compared with Kellgren-Lawrence and Ahlbäck ordinal grading: an exploratory study of knee radiographs using paired comparisons. *BMJ Open* 2021;**11**:e041793. doi:10.1136/bmjopen-2020-041793

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-041793).

Received 17 June 2020
Revised 23 December 2020
Accepted 21 January 2021



For numbered affiliations see end of article.

**Correspondence to**
Anders Odgaard;
anders.odgaard@regionh.dk

## ABSTRACT

**Objectives** Ordinal scales provide means for communicating the severity of a condition, but they are affected by cognitive biases, they introduce statistical problems and they sacrifice resolution. Clinicians discern more details than contained in scales, for example, when assessing radiographs, but clinicians' distinctions are often based on experience-based rules of thumb, that is, heuristics. The objectives of this study are to compare clinicians' heuristic assessments to ordinal grading, to identify case elements that influence clinicians' judgements and to present a method for quantifying heuristic assessments.

**Design** Clinicians were presented with 17 207 random pairs from a set of 1087 knee radiographs. For each pair, the radiograph with more severe osteoarthritis was selected. The Bradley-Terry model was used to calculate an osteoarthritis strength parameter for each radiograph. Similarly, strength parameters were determined for 12 morphological features with five additional features being considered either present or absent. All radiographs were also graded according to conventional ordinal systems (Kellgren-Lawrence and Ahlbäck). Relations between clinicians' judgements and (1) the heuristics-based osteoarthritis strength, (2) conventional ordinal systems and (3) morphological features were investigated.

**Results** Receiver operating characteristic analysis showed that the Bradley-Terry model provided a good description of clinicians' assessments (area under the curve (AUC)=0.97, 95% CI 0.968 to 0.972). Morphological features (AUC=0.90, 95% CI 0.900 to 0.908) provided a superior description of clinicians' choices compared with conventional ordinal systems (AUC=0.88, 95% CI 0.878 to 0.887 and AUC=0.80, 95% CI 0.796 to 0.809) for Ahlbäck and Kellgren-Lawrence, respectively). The features most strongly associated with osteoarthritis strength were medial joint space width, flattening of the medial femoral and tibial condyles, medial osteophytes and alignment.

**Conclusions** Heuristics-based assessments give a better distinction than conventional grading systems of knee osteoarthritis. The example presents a general approach to evaluate which features are part of experts' heuristics. The data suggest that experts discern more details than included in conventional ordinal grading systems.

## Strengths and limitations of this study

- This is the first study to demonstrate that clinicians consistently discern more details than contained in the acclaimed Kellgren-Lawrence and Ahlbäck ordinal scales.
- Using the Bradley-Terry model on clinicians' heuristics-based pairwise comparisons allows the calculation of a loss-less, ratio-scale strength or severity parameter for individual cases alleviating the problems of ordinal scales.
- The granularity and detail obtained with the described method is well suited for research questions where differences between radiographs or other qualitative information are studied.
- A unique understanding of the influence of underlying item features that determine clinicians' judgements can be achieved similarly using pairwise comparisons and the Bradley-Terry model.
- The study focused on osteoarthritic knee radiographs as an example, but the findings may reasonably be extended to other fields.

Quantitative heuristic assessments may replace ordinal scales.

## INTRODUCTION

Ordinal scales are used to quantify an underlying variable, for example, severity of a condition. Ordinal scales may be informal, for example, 'This is a bad case', or formal with class descriptions, for example, 'NYHA (New York Heart Association) heart failure class III'. Both formal and informal scales allow easy communication between clinicians, and they provide personal references for the individual clinician. Ordinal scales reduce continuous variation to categories with consequent loss of detail and statistical problems,[1] but there are further issues: they

are often based on expert opinion with post hoc validation, they often use unsubstantiated multidimensional class definitions, and they are affected by cognitive biases resulting in high intrarater and inter-rater variability.[2–4] Also, formal grading systems may prioritise concept over percept, that is, causing an observer to focus on what one expects to observe rather than what is actually available.[5]

Clinicians acquire experience through exposure to large volumes of patients and disease courses, and they interiorise relations between perceptions and clinical entities.[6] According to Polanyi, they develop tacit knowledge, that is, a deeper understanding that often cannot be explicitly stated: 'We can know more than we can tell'.[6] In 21st century terms, clinicians unknowingly develop algorithms and weighing factors analogous to machine learning. Increasing levels of tacit knowledge seem to promote intuitive judgements and data-driven rules of thumb.[7–10]

Methods, rules and strategies based on experience have traditionally been called 'heuristic'.[11] Developments in cognitive psychology[12–14] resulted in introduction of the noun 'heuristic', which may be defined as an efficient mental tool, a 'rule of thumb'. We will throughout this paper use the term 'heuristic' in this sense: 'a simple procedure that helps find adequate, though often imperfect, answers to difficult questions'.[14] Heuristics cause cognitive biases,[13–15] but they also offer advantages in decision making.[16–18]

The motivation for this study was frustration over the inability of ordinal scales to catch the levels of detail perceived by clinicians. It became apparent that clinicians can reproducibly discern differences in the severity of radiographic osteoarthritis (OA), even when radiographs are identically classified using acclaimed ordinal scales. Yet, clinicians may have difficulty in formulating discerning rules, and they often regress to 'obvious from experience'. The concepts of tacit knowledge and heuristics provide a model for studying, how clinicians make judgements. Case features, that is, perception elements, observed by clinicians correspond to particulate details,[6] and the features are knowingly or unknowingly used by clinicians when selecting a heuristic from their adaptive toolbox.[17] We aimed to investigate, how clinicians' heuristic assessments compare to ordinal grading. We also intended to study how case features affect heuristics assessments and thereby get an understanding of their tacit knowledge.

We present a method for calculating a strength or severity parameter for clinical cases based on clinicians' heuristic assessments, and the strength parameter offers ratio scale properties without loss of resolution and alleviating the statistical problems of ordinal scales.[1] We also demonstrate a method for investigating, how underlying case features influence the overall case assessment. The relatively simple task of judging the severity of radiographic knee OA is used as an example of a general class of heuristic judgements.

## METHODS

The proposed general method for quantifying heuristic assessments is based on pairwise comparisons of items (here knee radiographs). Pairwise comparisons are performed regarding an overall property (OA strength) and regarding identified feature variables (morphological features) that are likely to drive the overall property. The results of the pairwise comparisons are represented in a benchmark model (the Bradley-Terry model (BT)) and item parameters are extracted for both the overall property and the feature variables. Next, a model class should be chosen to analyse the results, and we chose a linear statistical model with the overall property as response variable (more generalised models and machine learning could also be used). Feature variable selection is performed by fitting models within the model class and comparing these, and we performed the selection using coefficients of determination. Finally, the proposed model is compared with the benchmark model in order to quantify how well the pairwise heuristics-based comparisons are described by the proposed model, and we used receiver operating characteristic (ROC) curves and area under the curve (AUC) scores. The individual steps will be explained below.

### Material

We used 1087 weight-bearing posteroanterior (PA) preoperative knee radiographs from patients listed for knee arthroplasty from an ongoing study, which is a prospective observational cohort study of 1452 consecutive primary knee arthroplasty patients, who were included from three high-volume centres from September 2016 to December 2017 with the aim of explaining consistent variation in revision rates among regions in Denmark.[19 20] Patients were included in the main study irrespective of diagnosis, aetiology, the degree of radiographic change and planned type of implant. The study group had no authority over the participating centres, and indications for surgery depended entirely on individual surgeons. Only radiographs from patients who were treated with a medial unicompartmental or total knee arthroplasty were included in the present study. Radiographs with primarily lateral OA on the PA view were excluded, and the radiographs were all of the operated knee. Some of the PA radiographs were near-normal, and we assume that other diagnostic modalities, for example, MRI or arthroscopy, would have indicated surgery in these cases, but it is also possible that a few patients had a total knee replacement for isolated patellofemoral OA. All left knee radiographs were horizontally mirrored to present all radiographs as right knee radiographs.

For all radiographs, we determined the degree of OA, that is, OA strength, based on heuristic comparisons of pairs of radiographs, and morphological features were determined similarly. All radiographs were also classified using the ordinal rating systems of Kellgren-Lawrence[21] (KL) and Ahlbäck.[22]

### Statistics of pairwise comparisons

The BT model[23 24] is a standard method for estimating an underlying continuous variable, for example, the degree of OA, from pairwise comparisons of items. It is useful in situations where the variable is a qualitative property. The estimation is done by parametrising the probability $P_{ab}$ that one item *a* is preferred over another item *b* regarding the property/variable. Because the BT model is usually parametrised using ratios, $P_{ab}$ is customarily reported on a logarithmic scale: $\log(P_{ab}/P_{ba}) = \beta_a - \beta_b$. For the estimation of $\beta$ for each item taken from N items, a constraint is needed: $\sum\beta=0$.

In the original description of the model, a nomenclature derived from sports was used. A comparison was termed a game, an assessor was a referee, an item was a player, the more successful player was termed the winner, and the lesser successful was the loser.

The BT model can be fitted using either Bayesian methods (maximum a posteriori estimation (MAP)) or maximum likelihood estimation. The MAP approach was used throughout this analysis. All analyses in this study were performed with R using the BradleyTerryScalable package.[25 26]

### Heuristic assessments of OA

From the sample of N radiographs, three times N/2 pairs were randomly drawn from the N!/(2(N-2)!) two-combinations, so that each radiograph was compared with three other radiographs. A web-based system was developed for the comparisons (Procordo, Copenhagen). Each image pair was displayed to an assessor, who would then select the radiograph with the more severe OA (figure 1A).

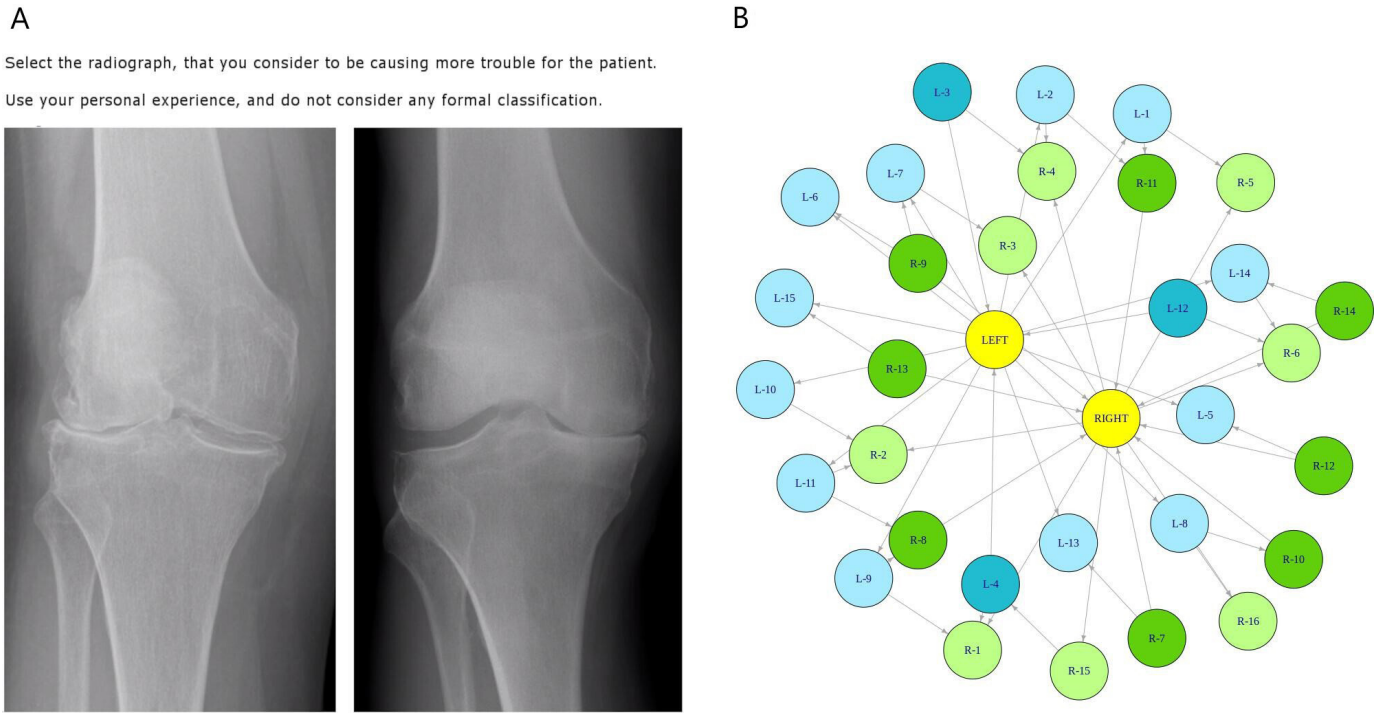


**Figure 1** A random pair or radiographs and their vicinity comparison network. (A) Display of a random pair of radiographs, from which the assessor had to choose the more severe. Severity was based on the assessor's personal experience of the relation between radiographic changes and severity of the patients' symptoms. The assessor was explicitly told not to consider any formal classification system. (B) The connectivity of comparisons in the vicinity of the two radiographs in (A) (marked as yellow vertices) with a maximum path length of three edges between the two yellow vertices. All radiographs in the vicinity network that were compared directly to the left radiograph are shown as blue vertices, and all radiographs that were compared directly to the right radiograph are shown as green vertices. Each edge between two vertices has an arrow that indicates the loser of the comparison, and an example is that radiograph 'L-3' (at the top of the graph) was found to be with more severe changes than radiograph 'R-4' (the loser). The colour shades of the green and blue vertices signify whether a radiograph was found to have more or less severe degenerative changes compared with the yellow vertex that they connect to. An example of a path from the left to the right radiograph is from 'left' to 'L-1' to 'R-11' to 'right', where all arrows indicate decreasing osteoarthritis strength. There are, however, also observations that could suggest that the right radiograph has more severe changes than the left, and an example is 'right' to 'R-15' to 'L-4' to 'left'. The total number of edges in the vicinity network pointing directly to 'left' is 3 (from the dark blue vertices), and the number pointing to 'right' is 8 (from the dark green vertices), which may be seen as an indication that the left radiograph has more degenerative changes (fewer losses) than the right. it should be noted, however, that the shown vicinity network is but a tiny fraction of the entire comparison network of 1087 vertices and 17 207 edges used to calculate the osteoarthritis strength of each radiograph. The two shown radiographs were, for instance, each compared with 33 other radiographs, but only the comparisons with a path between left and right of at most three edges are shown.

Thirteen experienced orthopaedic surgeons, who had been performing knee surgery (eleven exclusively, two also performed hip surgery) for at least 10 years, took the role of assessors. The surgeons were asked to select the radiograph from each pair that they considered to be causing more trouble for the patient based on their experience, and they were explicitly instructed not to consider any formal grading, but only to base judgements on clinical experience and intuition. No selection was possible for the first three seconds after display to reduce the effect of automaticity, and ties were not possible. The raw data output is a list of pairwise comparisons marking the more severe case.

Each radiograph was to be compared with 39 randomly selected radiographs, that is, three comparisons per surgeon, but a few of the surgeons did not complete their assignment, resulting in a total of 17207 comparisons. Thus, each radiograph was on average compared with 31.7 other radiographs (figure 1B). By use of the BT model, an osteoarthritis strength parameter $\beta_{OA}$ was assigned to each radiograph (figure 2A,B).

### Morphological features

We identified morphological features commonly reported when describing knee radiographs. The list of features was discussed among the participating surgeons and radiologists, which resulted in some features being added. All features were provisionally used to compare radiographs. Some features were found to be poorly defined or uncertain for use, and they were removed from the feature list. The result was a list of 17 morphological features (table 1), and we wanted to determine a set of feature strength parameters for each radiograph.

We found that most morphological features could be determined in all radiographs, while other features could only reasonably be identified in some of the radiographs (last five rows of table 1), and these were treated as binary variables. Two assessors (AO and KBGM) independently reviewed every radiograph to determine if each of the five features was present. If at least one of the assessors found a feature to be present in a radiograph, then the binary variable was set to 1, otherwise 0.

For each of the non-binary morphological features, the two assessors performed pairwise comparisons of radiographs (again, three comparisons per radiograph per assessor), with the aim of selecting the radiograph from each comparison, where the feature was more strongly present, and morphological feature parameters ($\beta_{JSW}$, $\beta_{LJS}$, $\beta_{MLS}$ etc.) were subsequently calculated for each radiograph using the BT model. Figure 2C,D shows $\beta_{MLS}$ and sample radiographs.

### Ordinal grading

The KL[21 27] and Ahlbäck[22] ordinal systems were used.[28]

The precise definition of the KL system is not explicitly stated in the original publication. For the purpose of this paper, the definition in radiopedia.org[29] was used (table 2). The grading is multidimensional, as both joint space narrowing, osteophytes, sclerosis and bone deformity are considered. The implicit assumption that these morphological features are monotonously related has never been substantiated.

The definition of the Ahlbäck classification in radiopedia.org[30] was used, adding a grade 0 for no OA (table 2). It may be argued that the morphological features of joint space narrowing and bone attrition are the results of a linear wear process, which would make the Ahlbäck grading unidimensional.

The grading according to both formal systems was done independently in two separate sessions by two musculoskeletal radiologists. First, each radiologist classified the radiographs in random order. For each presentation, the radiologist had to wait 3s before being allowed to grade the image in order to avoid hasty selections (Procordo, Copenhagen). All cases of disagreement were solved by consensus between the radiologists.

### Other statistics

The present study is an exploratory study, and no power considerations were relevant. Several linear models were created with the OA strength parameter $\beta_{OA}$ as response variable. Ordinal gradings were included as explanatory factor variables while morphological features were included as either continuous or factor (in the binary cases) explanatory variables. For the relevant linear models, ROC curves were created. For each comparison of two radiographs, the outcome of the comparison was used as response variable and the difference in the fitted values for each radiograph from the model was used as predictor. The area under each ROC curve (AUC) was reported.

The manuscript was prepared using the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist for cross-sectional studies.

### Patient and public involvement

No patients were involved in this study.

## RESULTS

### Heuristic assessment

The OA strength parameter $\beta_{OA}$ for the 1087 radiographs had a range from −14.8 to 5.9, with a mean of 0 as determined by the constraint $\Sigma\beta_{OA} = 0$ (figure 2A). Five examples of radiographs taken from the range of $\beta_{OA}$ are shown in figure 2B.

### Case features

The ranges of individual feature strength parameters are shown in table 1, and all were constrained by a mean of 0. Figure 2C,D shows $\beta_{MLS}$ for mediolateral subluxation. Several morphological features were mutually correlated, for example, ‘medial joint space width’ (JSW) and ‘alignment’ (ALG) (figure 3).

### Ordinal grading

Exact agreement was found in 71% and 59% of cases for KL and Ahlbäck, respectively. The distribution of KL and Ahlbäck grades after consensus is shown in table 2.

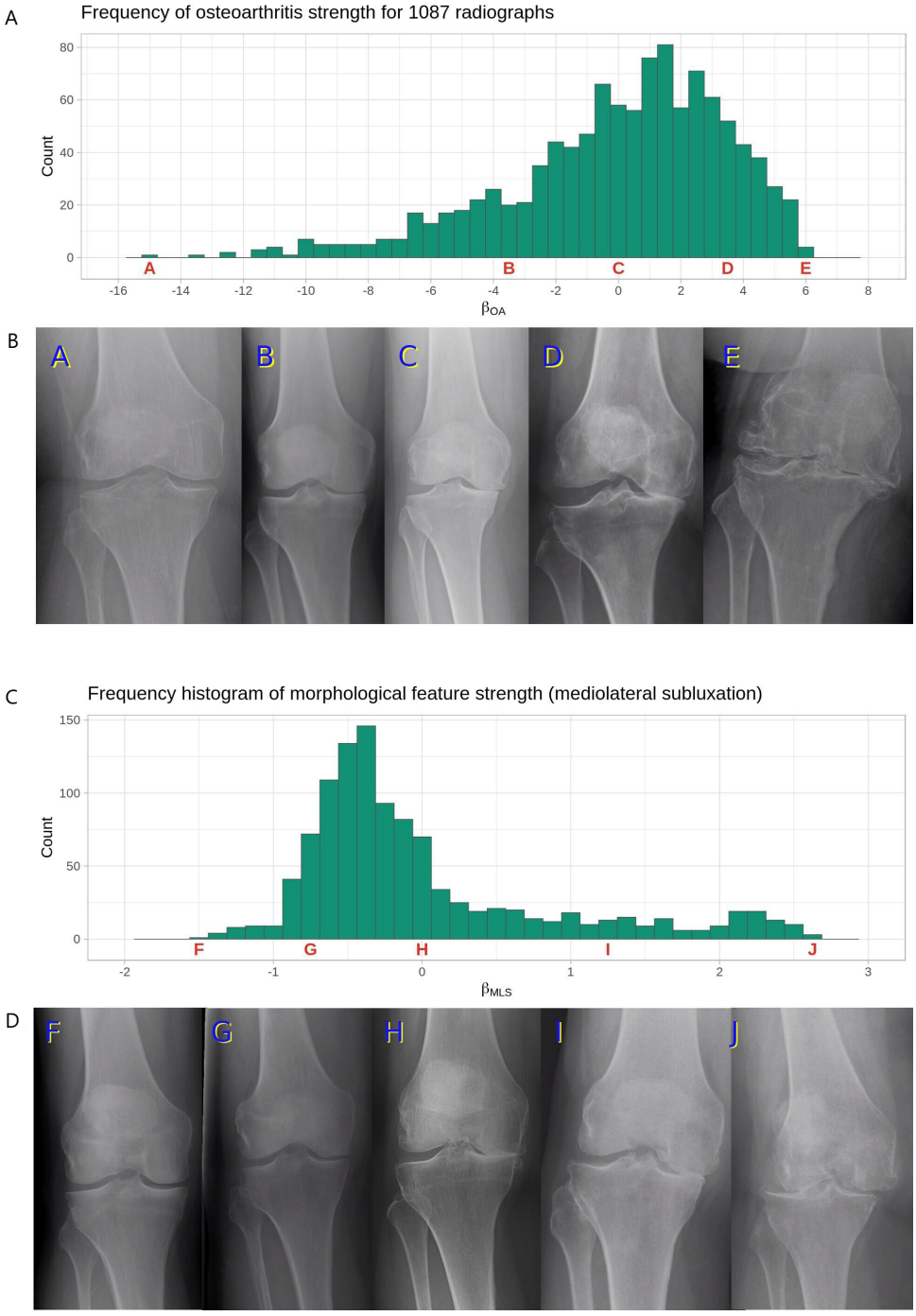


**Figure 2** These figures demonstrate the relation between the Bradley-Terry derived parameters and radiographs. (A) the frequency histogram of osteoarthritis (OA) strength parameter $\beta_{OA}$ of the 1087 radiographs studied. It should be noted, that $\beta_{OA}$ presents the degree of OA on a continuous scale rather than in a limited number of categories. A–E denote the $\beta_{OA}$ of the radiographs shown in (B). The radiographs A–E have been chosen to show the span of changes observed. (C) and (D) similarly show the span F–J of mediolateral tibiofemoral subluxation (MLS) observed in the dataset. MLS is but an example of the 17 different morphological parameters studied.

### Determinants of heuristic assessments

Linear relations between $\beta_{OA}$ and the morphological feature parameters were explored. Based on the sizes of correlation coefficients, the six features most strongly associated with $\beta_{OA}$ were (1) medial joint space width, (2) femur flattening, (3) tibia flattening, (4) medial osteophytes, (5) ALG and (6) wear groves (figure 3).

Performing a forward selection or a backward elimination of all morphological features in a linear model, with $\beta_{OA}$ as the response variable, resulted in the same model. All morphological variables except previous operation, pointing of tibial spines and chondrocalcinosis were included in the model, with ALG included as borderline significant with a p of 0.04 (when compared with

Table 1 List of morphological features included in the analyses

| Abbreviation | Name | Comments | β range (min; max) |
|---|---|---|---|
| JSW | Medial joint space width | 0-limited | –7.03 ; 2.19 |
| LJS | Lateral joint space width | 0-limited | –6.52 ; 3.55 |
| MLS | Mediolateral subluxation | Direction not specified. | –1.96 ; 2.74 |
| PAC | Periarticular calcification | | –2.07 ; 2.63 |
| OPM | Medial osteophytes | | –6.45 ; 4.08 |
| PTS | Pointing of tibial spines | | –7.45 ; 3.90 |
| OPL | Lateral osteophytes | | –7.23 ; 4.14 |
| ALG | Alignment | Dev. from normal. Direction not specified. | –6.44 ; 3.58 |
| FFL | Femur flattening | Medial condyle | –7.73 ; 3.82 |
| TFL | Tibia flattening | Medial condyle | –8.31 ; 4.07 |
| CYS | Cysts | | –2.23 ; 2.49 |
| SST | Subchondral sclerosis of tibia | Medial condyle | –3.98 ; 2.88 |
| CCN | Chondrocalcinosis | Feature only available in some radiographs | Binary 0/1 |
| WGR | Wear groves in the medial compartment | Feature only available in some radiographs | Binary 0/1 |
| AVN | Avascular necrosis | Feature only available in some radiographs | Binary 0/1 |
| MET | Metal implant | Feature only available in some radiographs | Binary 0/1 |
| POP | Bony signs of previous operation | Feature only available in some radiographs | Binary 0/1 |

Table 2 Definitions of the ordinal grades of Kellgren-Lawrence and Ahlbäck used in this paper

| Grade | Description | Distribution (%) |
|---|---|---|
| Kellgren-Lawrence ordinal grading | | |
| 0 | No radiographic features of OA are present | 7 (0.6) |
| 1 | Doubtful joint space narrowing (JSN) and possible osteophytic lipping | 62 (5.7) |
| 2 | Definite osteophytes and possible JSN on AP weight-bearing radiograph | 145 (13.3) |
| 3 | Multiple osteophytes, definite JSN, sclerosis, possible bony deformity | 807 (74.2) |
| 4 | Large osteophytes, marked JSN, severe sclerosis and definite bony deformity | 66 (6.1) |
| Ahlbäck ordinal grading | | |
| 0 | No radiographic features of OA are present | 60 (5.5) |
| 1 | Joint space narrowing (less than 3 mm) | 303 (27.9) |
| 2 | Joint space obliteration | 413 (38.0) |
| 3 | Minor bone attrition (0–5 mm) | 296 (27.2) |
| 4 | Moderate bone attrition (5–10 mm) | 12 (1.1) |
| 5 | Severe bone attrition (more than 10 mm) | 3 (0.3) |

The last column shows the distribution of the gradings in the 1087 radiographs after consensus.

the model without ALG through an analysis of variance test).

Three linear models were created: $\beta_{OA}$ as a response of Ahlbäck or KL grades and as a response of the feature

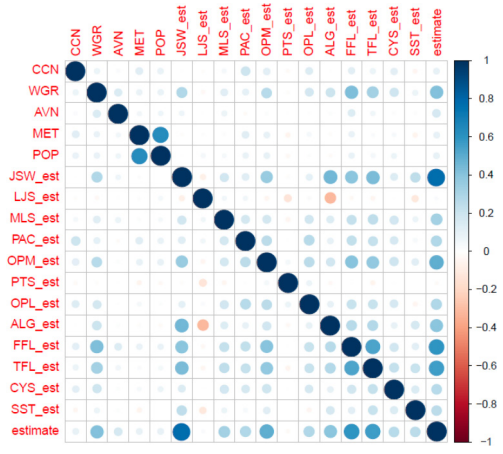


Figure 3 Correlation matrix showing associations between morphological features and the overall osteoarthritis (OA) strength estimate $\beta_{OA}$ from the Bradley-Terry model. Medial joint line space width (JSW) is for instance moderately correlated to alignment (ALG) but not to previous operation (POP). ALG, alignment; AVN, avascular necrosis; CCN, chondrocalcinosis; CYS, cysts; FFL, femur flattening; LJS, lateral joint space width; MET, metal implant; MLS, mediolateral subluxation; OPL, lateral osteophytes; OPM, medial osteophytes; PAC, periarticular calcification; PTS, pointing of tibial spines; SST, subchondral sclerosis of tibia; TFL, tibia flattening; WGR, wear groves in the medial compartment.

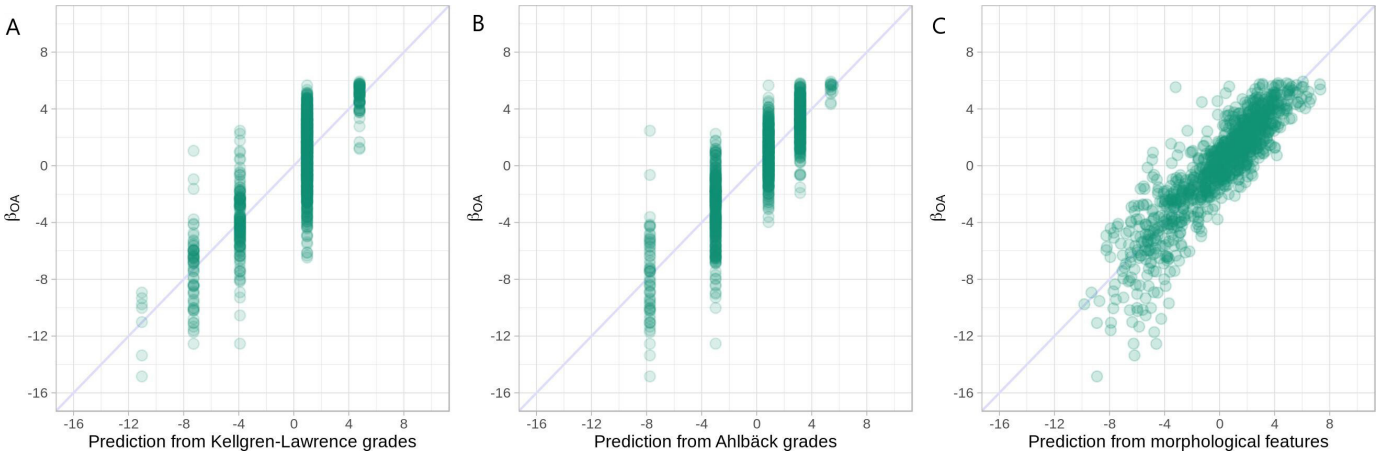


**Figure 4** The observed versus predicted osteoarthritis (OA) strength parameter $\beta_{OA}$ based on the three models (A) Kellgren-Lawrence ($r^2$=0.63). (B) Ahlbäck ($r^2$=0.73). (C) morphological features ($r^2$=0.79). Of the three models, the feature model explained the highest amount of variation in $\beta_{OA}$ followed by the Ahlbäck model and then the Kellgren-Lawrence model.

model. An indicator of how much variance of $\beta_{OA}$ was explained by each model is the adjusted $r^2$ ($\beta_{OA}$ ~ KL: $r^2$=0.63; $\beta_{OA}$ ~ Ahlbäck: $r^2$=0.73; $\beta_{OA}$ ~ $\sum\beta_{feature}$: $r^2$=0.79). Observed vs predicted values for the three models are shown in figure 4A–C. The feature model explained more variance in $\beta_{OA}$ than the models depending on either of the ordinal systems.

Seventy-three per cent of the variance in $\beta_{OA}$ could be modelled by a linear combination of (1) medial joint space width, (2) flattening of the medial femoral condyle, (3) flattening of the medial tibial condyle, (4) medial osteophytes and (5) alignment. Adding more morphological features resulted in only a modest explanatory improvement.

The AUCs for different ROC curves were calculated to compare different predictions of the experts' heuristic decisions (figure 5). The AUC for the surgeons' decisions predicted by $\beta_{OA}$ was 0.97 (95% CI 0.968 to 0.972) indicating that the BT model provides a good description of data. The AUC for predictions from the Ahlbäck or KL score was 0.88 (95% CI 0.878 to 0.887) and 0.80 (95% CI 0.796 to 0.809), respectively. The AUC for predictions from the feature model was 0.90 (95% CI 0.900 to 0.908). Note, that the CIs for the AUC scores are non-overlapping. This shows that the feature model was able to explain the surgeon's decisions more precisely than the ordinal grades (Ahlbäck better than KL).

## DISCUSSION

This paper describes a general method for quantifying clinical heuristic assessments, that is, based on tacit knowledge, and for revealing the individual case features that matter to clinicians. By using the BT model, an underlying variable, for example, the strength of a clinical observation, can be expressed on a ratio scale.[31] For the example studied, we found that the BT model provides a good description of clinicians' choices, and we have shown that experts discern greater levels of detail than provided by acclaimed ordinal scales. By relating the strength of individual features to the overall case strength, it is possible to demonstrate which morphological features—possibly unknowingly—matter to experts.

For the particular example of knee OA, it was found that a linear model including the morphological features of medial joint space width, flattening of the medial

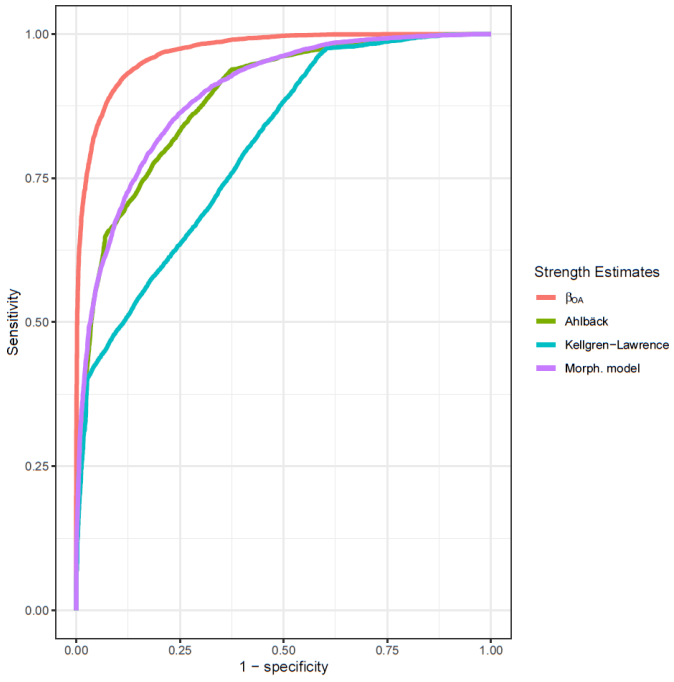


**Figure 5** Receiver operating characteristic curves for different predictions of the experts' heuristic decisions. The AUC for the surgeons' decisions predicted by $\beta_{OA}$, Ahlbäck, Kellgren-Lawrence and the morphological feature model was 0.97 (95% CI 0.968 to 0.972), 0.88 (95% CI 0.878 to 0.887), 0.80 (95% CI 0.796 to 0.809) and 0.90 (95% CI 0.900 to 0.908), respectively. The morphological feature model was better than any of the ordinal grading models, and the Ahlbäck model was better than the Kellgren-Lawrence model in predicting the surgeons' heuristic decisions when comparing random radiographs. AUC, area under the curve; OA, osteoarthritis.

femoral condyle, flattening of the tibial condyle, medial osteophytes and alignment provided a good description of the clinicians' heuristics-based assessments, and we conclude that these morphological features are important to clinicians. It should be noted that the first four of these features are included in either the KL or the Ahlbäck ordinal systems, but they are not included in both systems, and none of the systems consider alignment or some of the less important features, that matter to clinicians.

Many ordinal grading systems are the result of originators' heuristics and intuitive grading. Our method can be viewed as a way of validating the level of tacit knowledge of grading systems' originators. Alternatively, observed relations between an overall strength parameter and individual feature strength parameters based on many experts' assessments may provide more robust designs of ordinal grading systems.

Our data set consisted of radiographs of knees listed for arthroplasty, which imposes a limitation on the conclusions regarding the relation between the overall OA strength parameter and the morphological features. The regression results should not be extrapolated to other settings, for example, primary care, since only a few cases with very mild OA were included. The fact that some of the morphological features were non-linearly related to the overall strength parameter also suggests that the relations between features and the overall strength parameter could be studied in more detail. The main conclusions of the study, that clinicians' heuristic assessments discern more detail than contained in the classic ordinal scales, and that the method allows the examination of relations between an overall strength parameter and item features are, however, not affected by this limitation.

The purpose of the paper was to present a method for quantifying heuristic assessments and as such we have not performed sensitivity analyses or validation of the developed linear models and the achieved results. Validation may be done using different methods, for example, k-fold cross-validation, to quantify the importance of the specific data on the variables included in the final model and the estimates for these variables. Such a way of determining the variability of the data-driven model could be included in our proposed general method. It is an important point that if the approach proposed in this paper is to be used to identify case features (or other decision-driving variables), validation and sensitivity analyses of the identification process and the final variables included should be performed.

It is the assumption of our approach, that experienced clinicians possess tacit knowledge.[6] They are assumed to have collected experience and observed relations through years of practice, that have become internalised and may yet be difficult to express. Our method provides a way for detailing the particulars, for example, morphological features, used by experienced clinicians when judging a comprehensive entity, for example, a radiograph. By explicitly integrating the particulars, our understanding of clinician judgement is expanded, and we may gain important insights. The validity of tacit knowledge has, however, not been our focus. By relating the heuristics-based strengths of items to some external observation, for example, patient-reported outcome, external validity of the heuristic assessment may be demonstrated.

The Bradley-Terry model has rarely been applied in medical research. In 1995 Matthews and Morris used the model in a study of pain.[32] Searching for BT in PubMed yields 36 results (search of 23rd Feb. 2020), but none of the papers are related to clinical research.

Classification of images is an obvious field of application for the method. There are clinical situations where classifications have a dichotomous character and a multilevel grading is not needed, for example, when classifying lumbar stenosis, but we suggest that the method should be used in studies where detail and discriminative power is needed to distinguish grades of indications, outcomes, etc.

The low reliability of many ordinal grading systems may be explained by a multitude of cognitive biases.[14 33–35] It is plausible that methods based on pairwise comparisons will reduce some of this bias, but further research is needed to substantiate this assumption. We have shown that experts are able to distinguish between grades of OA severity not accessible by the grades of the conventional ordinal systems, which may well be important for decision making and research. Based on our findings, many studies claiming (lack of) relations between qualitative data and other observations may be questioned.

We used a simple linear model to relate all morphological features to overall OA strength. A more sophisticated way of performing variable selection is to consider intervariable correlations along with plots of marginal distributions. The marginal distributions showed that the effects of several of the morphological feature parameters were non-linear in nature, and some features were highly correlated.

The suggested method for assigning a severity strength and extracting elements of tacit knowledge is well suited for groups of cases. In situations, where a severity strength is to be assigned to a single case, this approach cannot be used. For use in single cases, we suggest that reference sets be established, which can be used for comparisons.

The purpose of this study was to describe methods for quantifying and investigating experts' heuristics-based assessments. We chose to investigate radiographs of osteoarthritic knees, but the methods may be used for any item type of data including complex information aggregating both quantitative and qualitative data.

**Author affiliations**
[1]Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark
[2]Department of Radiology, Copenhagen University Hospital Herlev-Gentofte, Copenhagen, Denmark
[3]Department of Orthopedic Surgery, Copenhagen University Hospital Herlev-Gentofte, Copenhagen, Denmark
[4]Department of Orthopedic Surgery, Rigshospitalet - Copenhagen, University Hospital, Copenhagen, Denmark

[5]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

**Acknowledgements** We thank radiologist Lone Rømer for performing ordinal grading and the following orthopedic surgeons for contributing to the heuristic assessments: Thomas Bjerno, Claus Fink Jepsen, Andreas Kappel, Mogens Laursen, Thomas Lind, Frank Madsen, Lars Peter Møller, Lasse Enkebølle Rasmussen, Søren Rytter, Henrik Schrøder, Snorre Stephensen and Svend Erik Østgaard. We thank the Capital Region of Denmark for supporting this study.

**Contributors** AO and KBC conceived and designed the study. AM-P collected the radiographs used. AO designed the software for the heuristic assessments. KBGM and AO performed heuristic assessments, and KBGM performed the ordinal grading. MMP, KBC and AO performed the statistical analyses. AO wrote the first and updated drafts of the article, with important contributions mainly from MMP and KBGM, while AM-P and KBC commented and approved.

**Funding** The Health Research Foundation of the Capital Region of Denmark funded this research (E-19255-50-02, granted 1 July 2015).

**Competing interests** AO is co-owner of Procordo Software that provided the software framework used for data collection. No other relationships or activities that could appear to have influenced the submitted work.

**Patient consent for publication** Not required.

**Ethics approval** The National Committee of Health Research Ethics provided ethical approval (Protocol no. 16038343, 2 September 2016) and The Danish Data Protection Agency (Jr. no. 2012-58-0004, HGH-2016–087, I-Suite no. 04819) approved data management.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Deidentified raw data with results of pairwise comparisons are available on reasonable request. Data are not available for online access. Readers who wish to gain access to the data can write to the senior author AO at anders.odgaard@regionh.dk with their requests.



## REFERENCES

1 Forrest M, Andersen B. Ordinal scale and statistics in medical research. *Br Med J* 1986;292:537–8.
2 Mak PHK, Campbell RCH, Irwin MG, *et al*. The ASA physical status classification: inter-observer consistency. American Society of Anesthesiologists. *Anaesth Intensive Care* 2002;30:633–40.
3 Grootenboer EMV, Giltay EJ, van der Lem R, *et al*. Reliability and validity of the global assessment of functioning scale in clinical outpatients with depressive disorders. *J Eval Clin Pract* 2012;18:502–7.
4 Claessen FMAP, Meijer DT, van den Bekerom MPJ, *et al*. Reliability of classification for post-traumatic ankle osteoarthritis. *Knee Surg Sports Traumatol Arthrosc* 2016;24:1332–7.
5 Bleakley A, Farrow R, Gould D, *et al*. Making sense of clinical Reasoning: judgement and the evidence of the senses. *Med Educ* 2003;37:544–52.
6 Polanyi M. *The Tacit dimension*. New York: Doubleday, 1966.
7 Bowen L, Shaw A, Lyttle MD, *et al*. The transition to clinical expert: enhanced decision making for children aged less than 5 years attending the paediatric ED with acute respiratory conditions. *Emerg Med J* 2017;34:76–81.
8 Stolper E, Van de Wiel M, Van Royen P, *et al*. Gut feelings as a third track in general practitioners' diagnostic reasoning. *J Gen Intern Med* 2011;26:197–203.
9 Norman G, Young M, Brooks L. Non-analytical models of clinical Reasoning: the role of experience. *Med Educ* 2007;41:1140–5.
10 Patel VL, Arocha JF, Kaufman DR, Review KDR. A primer on aspects of cognition for medical informatics. *J Am Med Inform Assoc* 2001;8:324–43.
11 Groner M, Groner R, Bischof WF. Approaches to heuristics: a historical review. In: Groner R, Groner M, Bischof WF, eds. *Methods of heuristics*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers, 1983: 1–18.
12 Simon HA, Newell A. Human problem solving: the state of the theory in 1970. *Am Psychol* 1971;26:145–59.
13 Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. *Science* 1974;185:1124–31.
14 Kahneman D. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux, 2011.
15 Itri JN, Patel SH. Heuristics and cognitive error in medical imaging. *AJR Am J Roentgenol* 2018;210:1097–105.
16 Gigerenzer G, Goldstein DG. Reasoning the fast and frugal way: models of bounded rationality. *Psychol Rev* 1996;103:650–69.
17 Gigerenzer G. Why Heuristics work. *Perspect Psychol Sci* 2008;3:20–9.
18 Goldstein DG, Gigerenzer G. Models of ecological rationality: the recognition heuristic. *Psychol Rev* 2002;109:75–90.
19 Mørup-Petersen A. *Evaluation of Danish knee replacement surgery: patient-reported outcomes versus register data*. Copenhagen: University of Copenhagen, 2020.
20 Mørup-Petersen A, Laursen M, Madsen F, *et al*. Large variation in revision rates after primary knee arthroplasty: a matter of patient selection? baseline data from 1452 patients in the prospective multicenter cohort study, spark. *Submitted*.
21 Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. *Ann Rheum Dis* 1957;16:494–502.
22 Ahlbäck S, Rydberg J. [X-ray classification and examination technics in gonarthrosis]. *Lakartidningen* 1980;77:2091–3.
23 Bradley RA, Terry ME. Rank analysis of incomplete block designs. 1. The method of paired comparisons. *Biometrika* 1952;39:324–45.
24 Agresti A. *Categorical data analysis*. Hoboken: John Wiley & Sons, 2014.
25 Kaye E, Firth D. Package 'BradleyTerryScalable', 2017. Available: https://github.com/EllaKaye/BradleyTerryScalable
26 R Core Team. A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria., 2020. Available: https://www.R-project.org/
27 Kohn MD, Sassoon AA, Fernando ND. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. *Clin Orthop Relat Res* 2016;474:1886–93.
28 Braun HJ, Gold GE. Diagnosis of osteoarthritis: imaging. *Bone* 2012;51:278–88.
29 Radiopaedia. Kellgren and Lawrence system for classification of osteoarthritis of knee, 2020. Available: https://radiopaedia.org/articles/kellgren-and-lawrence-system-for-classification-of-osteoarthritis-of-knee
30 Radiopaedia. Ahlbäck classification of osteoarthritis of the knee joint, 2020. Available: https://radiopaedia.org/articles/ahlback-classification-of-osteoarthritis-of-the-knee-joint
31 Stevens SS. On the theory of scales of measurement. *Science* 1946;103:677–80.
32 Matthews JNS, Morris KP. An application of Bradley-Terry-type models to the measurement of pain. *Appl Stat* 1995;44:243–55.
33 Aeffner F, Wilson K, Martin NT, *et al*. The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. *Arch Pathol Lab Med* 2017;141:1267–75.
34 Brady AP. Error and discrepancy in radiology: inevitable or avoidable? *Insights Imaging* 2017;8:171–82.
35 Croskerry P. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Acad Emerg Med* 2002;9:1184–204.