RMD Open

Rheumatic & Musculoskeletal Diseases

ORIGINAL RESEARCH

Development of radiographic classification criteria for hand osteoarthritis: a methodological report (Phase 2)

Ida K Haugen [ORCID],[1] David Felson [ORCID],[2,3] Abhishek Abhishek,[4,5] Francis Berenbaum [ORCID],[6,7] John James Edwards,[8] Gabriel Herrero Beaumont,[9] Merete Hermann-Eriksen,[1] Catherine L Hill,[10,11] Mariko Ishimori,[12] Helgi Jonsson,[13] Teemu Karjalainen,[14] Ying Ying Leung,[15] Emmanuel Maheu,[7] Christian D Mallen,[8] Rikke Helene Moe [ORCID],[16] Roberta Ramonda [ORCID],[17] Valentin Ritschl [ORCID],[18,19] Tanja A Stamm [ORCID],[18,19] Zoltan Szekanecz [ORCID],[20] Florus J van der Giesen,[21] Marco J P F Ritt,[22] Ruth Wittoek,[23] Ingvild Kjeken [ORCID],[16] Nina Osteras [ORCID],[16] Lotte A van de Stadt,[21] Martin Englund,[24] Krysia S Dziedzic,[8] M Marshall [ORCID],[8] Sita Bierma-Zeinstra,[25] Paul Hansen,[26] Elsie Greibrokk,[1] Wilma Smeets,[21] Margreet Kloppenburg[21,27]

For numbered affiliations see end of article.

Correspondence to
Dr Ida K Haugen;
ida.k.haugen@gmail.com

ABSTRACT

Objectives In Phase 1 of developing new hand osteoarthritis (OA) classification criteria, features associated with hand OA were identified in a population with hand complaints. Radiographic findings could better discriminate patients with hand OA and controls than clinical examination findings. The objective of Phase 2 was to achieve consensus on the features and their weights to be included in three radiographic criteria sets of overall hand OA, interphalangeal OA and thumb base OA.

Methods Multidisciplinary, international expert panels were convened. Patient vignettes were used to identify important features consistent with hand OA. A consensus-based decision analysis approach implemented using 1000minds software was applied to identify the most important features and their relative importance influencing the likelihood of symptoms being due to hand OA. Analyses were repeated for interphalangeal and thumb base OA. The reliability and validity of the proposed criteria sets were tested.

Results The experts agreed that the criteria sets should be applied in a population with pain, aching or stiffness in hand joint(s) not explained by another disease or acute injury. In this setting, five additional criteria were considered important: age, morning stiffness, radiographic osteophytes, radiographic joint space narrowing and concordance between symptoms and radiographic findings. The reliability and validity were very good.

Conclusion Radiographic features were considered critical when determining whether a patient had symptoms due to hand OA. The consensus-based decision analysis approach in Phase 2 complemented the data-driven results from Phase 1, which will form the basis of the final classification criteria sets.

Key messages

What is already known about this subject?
► New classification criteria sets, using radiographs and not clinical examination findings alone, are needed to classify overall hand osteoarthritis (OA), interphalangeal OA and thumb base OA.

What does this study add?
► The proposed criteria sets should be applied in a population with hand symptoms not explained by another disease or acute injury.
► The most important hand OA features and their relative weights were determined using a consensus-based decision analysis approach.
► In the proposed criteria sets, radiographic features and concordance between radiographic findings and symptoms were considered most important (ie, given highest weight). Important self-reported features included morning stiffness and age, which were included in the proposed criteria sets.

How might this impact on clinical practice or further developments?
► New radiographic classification criteria sets for hand OA and its subsets will enable more homogeneous inclusion of patients across both observational studies and clinical trials.

INTRODUCTION

Valid and reliable classification criteria sets for hand osteoarthritis (OA) and its subsets enable more homogeneous inclusion of patients in observational studies and clinical

trials. This paper describes parts of the process of developing new classification criteria sets for hand OA, which includes two separate steps. The first step leads to the development of classification criteria based on radiographic findings and symptoms, which will be labelled 'radiographic hand OA criteria'. The second step leads to clinically defined hand OA criteria. We aim to develop three criteria sets in each step: for overall hand OA, interphalangeal OA and thumb base OA. These latter two subsets represent distinct phenotypes with differences in pathogenesis and treatment.[1]

The development of the radiographic criteria sets involves three phases. In Phase 1, we collected self-reported, clinical, laboratory and radiographic data on patients with and without hand OA from primary and secondary/tertiary care centres. We identified features that could differentiate patients with and without hand OA, which could serve as key factors in the new criteria sets.[2] In Phase 2, using decision-making software, the aim was to determine the relative weight of factors that influence the expert opinions about the probability of a person having symptoms due to hand OA. This approach allowed us to capture the experts' clinical perspectives, which supplement the data-driven results from Phase 1. Results from these first two phases will be integrated to form the final sets of radiographic criteria in Phase 3. The current paper outlines the details of Phase 2 in identifying and weighing the factors that experts attributed to the radiographic classification criteria sets for overall hand OA, interphalangeal OA and thumb base OA.

METHODS

Two panels of multidisciplinary, international experts were convened. Several sets of patient vignettes representing patients with hand symptoms and findings were created based on information about actual patients from Phase 1.[2] We applied a consensus-based decision analysis approach implemented using 1000minds software (www.1000minds.com) to identify the most important criteria and weights representing their relative importance influencing the likelihood of symptoms being due to hand OA (figure 1). 1000minds software has been widely used in similar efforts for developing disease classification criteria sets in rheumatology.[3–7] The expert panel participated in three 1000minds surveys. Finally, the reliability and validity of the criteria and weights were tested. The approach, as explained in more detail below, was first completed for overall hand OA before parts of it were repeated for interphalangeal OA and thumb base OA, respectively.

The expert panels

Two expert panels with different compositions and tasks were assembled. The 1000minds surveys were performed by Panel 1, including 21 experts with multidisciplinary background who actively see patients with hand OA in their clinic and have academic experience in hand OA research. Panel 1 was involved in surveys where clinical experience was considered important. The reliability exercise was performed by Panel 2, including seven experts with experience in OA research. Four of these seven experts were actively involved in the project from its beginning, whereas the other three were unfamiliar with the project. By including experts who were unfamiliar with the project, the results on reliability were less biased.

The 1000minds surveys were designed and administered by IKH in collaboration with the chair (MK), the methodologist (DF) and a co-inventor of 1000minds (PH).
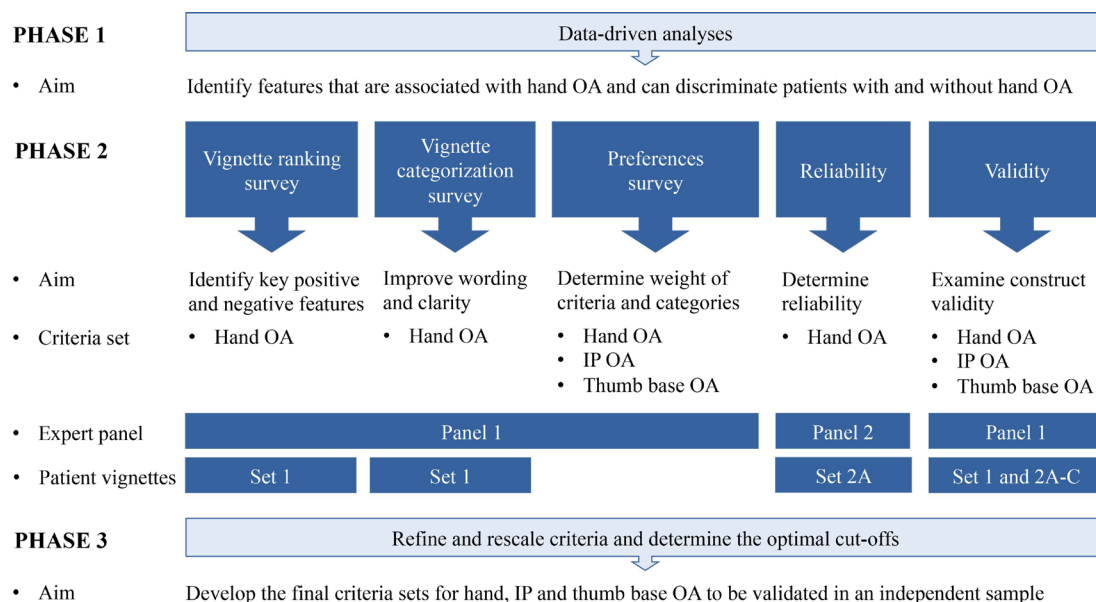

PHASE 1	Data-driven analyses				
• Aim	Identify features that are associated with hand OA and can discriminate patients with and without hand OA				
PHASE 2	Vignette ranking survey	Vignette categorization survey	Preferences survey	Reliability	Validity
• Aim	Identify key positive and negative features	Improve wording and clarity	Determine weight of criteria and categories	Determine reliability	Examine construct validity
• Criteria set	• Hand OA	• Hand OA	• Hand OA • IP OA • Thumb base OA	• Hand OA	• Hand OA • IP OA • Thumb base OA
• Expert panel	Panel 1			Panel 2	Panel 1
• Patient vignettes	Set 1	Set 1		Set 2A	Set 1 and 2A-C
PHASE 3	Refine and rescale criteria and determine the optimal cut-offs				
• Aim	Develop the final criteria sets for hand, IP and thumb base OA to be validated in an independent sample				

Figure 1 Overview of the development of classification criteria sets for overall hand OA, interphalangeal OA and thumb base OA. IP, interphalangeal; OA, osteoarthritis.

Development of patient vignettes

Each patient vignette was written as a short story describing a real-life patient from Phase 1.[2] To maintain anonymity, each case was assigned a fictitious name.

Several sets of patient vignettes were created. In the first set of 25 patient vignettes (Set 1), we included a mixture of patients with their symptoms clearly due to hand OA, clearly due to another disease and patients whose symptoms were of unclear cause. Each vignette had four sections: (1) an introduction about demographic factors and symptoms, (2) clinical examination findings, (3) laboratory results and (4) hand radiography results (online supplemental file S1). These patient vignettes were used in the first two 1000minds surveys (vignette ranking and vignette categorisation surveys).

Three separate sets (Set 2A–C), each including 30 patient vignettes, were developed after the 1000minds surveys to test the reliability and validity of the three proposed criteria sets. Since these vignettes should be used to test the validity of the three criteria sets in the target population, symptoms in the relevant joint group had to be present and no other disease or acute injury should be present to explain the symptoms in these patients. Set 2A included patients with symptoms in interphalangeal and/or thumb base joints, and was used in the surveys for overall hand OA. Two additional sets of patient vignettes were subsequently developed, including patients with interphalangeal joint symptoms (Set 2B) and thumb base symptoms (Set 2C), respectively. Importantly, there was no overlap between Set 1 and Sets 2A–C in order to avoid that validity was tested in the same population that was used for identification of key features. The patient vignettes in Sets 2A–C included information about the proposed criteria only (online supplemental file S1).

Vignette ranking surveys (individual surveys and consensus survey)

Following a review of the results from Phase 1, each expert in Panel 1 individually participated in an online 1000minds survey, where they were asked to rank 25 patient vignettes (Set 1) according to how likely it was that the patient presented with hand OA (1st=most likely and 25th=least likely) (ie, individual surveys). The experts were informed that many patients could potentially have OA in their joints, which was not necessarily the cause of their symptoms. Hence, the experts should consider all available information, including demographic factors, clinical, laboratory and imaging findings, joint distribution of symptomatic joints, whether the symptoms occurred in joints with OA features and symptom characteristics. The mean and distribution of expert panel rankings for each case were plotted, and inter-rater agreement was assessed by the Kendall's coefficient of concordance (0=no agreement and 1=complete agreement).

A 3-hour webinar was arranged for the expert panel members. The results from the individual survey were presented by IKH before the group ranked the patient vignettes according to their probability of having symptoms due to hand OA (ie, consensus survey). The mean ranking from individual surveys was used as a starting point, and two and two patients were thereafter compared and ranked. In-depth discussions and arguments about why one case should be ranked higher or lower than another case enabled the identification of key positive and negative factors related to symptomatic hand OA. Results from Phase 1 were used to support the discussion and the ranking when needed. The group ranking of patients later served as a pseudo-gold standard when testing the validity of the hand OA criteria set.

Identification of domains and categories

The project chairs (IKH and MK) and the methodologist (DF) proposed a target population in which the criteria sets should be applied and a list of hand joints that should be considered, that is, the target joints. Based on the consensus ranking survey discussion, important key factors for the classification of overall hand OA were identified. They also identified additional key factors missing in the initial patient vignettes that could facilitate the classification of overall hand OA.

Categories within each criterion were proposed using a data-driven approach with Phase 1 data together with the experts' clinical experience. The definitions of short versus long morning stiffness and low versus high levels of inflammatory biomarkers were based on calculations of area under the receiver operating curve (AUC). When defining the categories for the radiographic and symptomatic hand OA criteria, we calculated the number of affected hand joints per patient before calculating the proportion of patients with hand OA across the different numbers of affected joints.

The list of proposed criteria and their categories were presented for the expert panel members in a webinar. The experts agreed on a modified set of criteria and categories that should be tested in the following 1000minds categorisation survey.

For interphalangeal OA and thumb base OA criteria sets, the experts agreed on the same list of proposed criteria as for overall hand OA. Hence, the vignette case ranking surveys were not repeated for these two criteria sets. The categories for interphalangeal OA were left identical to the categories for overall hand OA. In contrast, the categories about the number of joints affected were changed for thumb base OA due to considerably smaller number of potentially affected joints than for the whole hand.

Vignette categorisation survey (individual surveys)

The proposed criteria set for overall hand OA was tested on our patient vignettes to detect possible problems with wording of the criteria and categories. Based on available information in the 25 patient vignettes (Set 1), the Panel 1 experts were instructed to choose the correct category for each criterion as shown in online supplemental file S2. A webinar was arranged to discuss the results from

QUESTION #1

Which of these two imaginary patients is more
likely to have pain aching or stiffness due to hand OA?

Assume the two patients are the same on all other criteria, at their "average" levels.
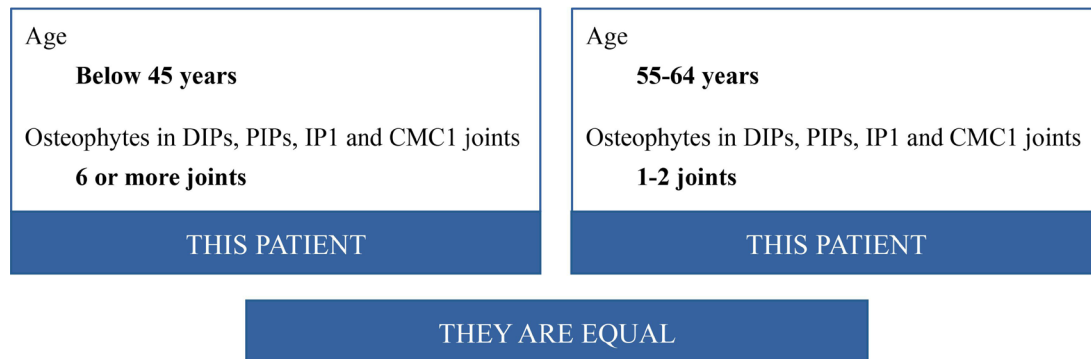
Age	Age
Below 45 years	**55-64 years**
Osteophytes in DIPs, PIPs, IP1 and CMC1 joints	Osteophytes in DIPs, PIPs, IP1 and CMC1 joints
6 or more joints	**1-2 joints**
THIS PATIENT	THIS PATIENT

THEY ARE EQUAL

Figure 2 Example of a trade-off question from the 1000minds preferences survey. CMC1, first carpometacarpal; DIP, distal interphalangeal; IP1, first interphalangeal; OA, osteoarthritis; PIP, proximal interphalangeal.

the vignette categorisation survey, and changes in the proposed criteria and their categories were made accordingly.

Preferences survey (pairwise comparison method—individual surveys)

In order to determine the weights for each criterion and category, representing their relative importance, the Panel 1 experts completed three 1000minds preferences surveys for overall hand OA, interphalangeal OA and thumb base OA, respectively. Instead of guessing the weights or assuming they are equally important, 1000minds determines them using the PAPRIKA method—an acronym for Potentially All Pairwise RanKings of all possible Alternatives.[8]

The method involved each expert being asked to answer, based on their clinical experience and judgement, a series of questions in terms of which patient's symptoms were more likely to be due to hand OA. Each question was based on choosing between two hypothetical patients defined by two criteria at a time and involving a trade-off. The experts were informed that the patients were otherwise the same, that is, they did not differ with respect to all other criteria. Figure 2 shows an example of a trade-off question where one patient is young but has many osteophytes, whereas the other patient is older but has fewer osteophytes (ie, involving a trade-off between age and osteophytes). Such questions (always involving a trade-off between two criteria at a time) were repeated with different pairs of hypothetical patients. Each time the expert answered a trade-off question, that is, ranked a pair of patients (including potentially ranking them equally), all other pairs of patients that could be pairwise ranked by applying the logical property of 'transitivity' were identified and eliminated by the software. For example, if an expert ranked patient X ahead of patient Y and patient Y ahead of patient Z, then, by transitivity, X

was also ranked ahead of Z (and was, therefore, not asked about by the software).

The 1000minds software uses mathematical methods to determine the weights for each category within each criterion.[8] Weights were calculated for each expert separately and averaged across all experts. The relative importance of each criterion is represented by the weight assigned to its highest-ranked category and the sum of these weights across the criteria is 100% (confirming each criterion's weight can be interpreted in relative terms). Thus, possible total scores for each patient range from 0 to 100, with higher scores indicating higher likelihood of the symptoms being due to hand OA.

The results from the preferences surveys were presented by IKH in webinars with the expert panel, with the weights and any possible redundancies of the criteria discussed.

Evaluation of validity

In the first validity exercise, the 25 patient vignettes in Set 1 were scored by IKH according to the criteria set for overall hand OA, resulting in a total score (range: 0–100) for each patient. The ranking of patients' total scores was compared with the consensus ranking (pseudo-gold standard), and the Spearman correlation coefficient was calculated. After removing possible redundant criteria, as decided by the experts, the patient vignettes were re-ranked according to their updated total scores, and the correlation coefficient was re-calculated.

In the second validity exercise, Panel 1 experts were asked to imagine that they would include patients in a clinical trial of a hypothetical promising disease-modifying drug that could halt OA progression and lead to less pain and stiffness. They were informed that all patients fulfilled the mandatory criteria about symptoms in at least one target joint on most days in the previous 6 weeks, and no other disease or acute injury explained the symptoms.

Based on the available information about the additional proposed criteria, they were asked to answer 'Yes' or 'No' as to whether they would include each of the 30 patients in the hypothetical clinical trial. Three separate surveys were done using Set 2A for overall hand OA, Set 2B for interphalangeal OA and Set 2C for thumb base OA. The patients in each set were ranked according to their total scores by IKH. The proportion of experts who wanted to include each of the 30 patients in the hypothetical trial was presented in a plot, and the Spearman correlation coefficient for the total score (range: 0–100) relative to the number of experts who included the patient in the trial (range: 0–21) was calculated.

Evaluation of reliability

The Panel 2 experts performed a reliability exercise for overall hand OA to detect inconsistencies in interpreting the criteria and categories. Using the 1000minds software, the experts read each patient vignette in Set 2A and chose the correct category for each criterion based on the available information in the patient vignette (online supplemental file S1). For each criterion, we calculated multirater free-marginal kappa values (http://justus-randolph.net/kappa/) and percentage agreement. In addition, intraclass correlation coefficients (ICC; mixed effect, absolute agreement and average measures) for the estimated total scores (range: 0–100) were computed.

Statistical analyses

The majority of analyses as explained in the sections above were automatically performed by the 1000minds software. When needed, additional analyses (eg, reliability and validity) were done using IBM SPSS Statistics V.26.

Patient and public involvement and engagement

Two patient partners (EG and WS) with lived experience of hand OA were involved in project design meetings, contributed in expert discussions and supported all phases of the study. They did not participate in the 1000minds surveys.

RESULTS

Panel 1 experts had a multidisciplinary background, including medical doctors within rheumatology (n=13), primary care (n=2) and surgery (n=2), occupational therapists (n=2), a physical therapist (n=1) and a physician assistant (n=1). The panel consisted of men (n=13) and women (n=8), and the experts were spread across Europe (n=17), North America (n=2), Asia (n=1) and Australia (n=1). In addition, two female patient partners from Norway and the Netherlands were involved. Panel 2 included medical doctors (n=2), occupational therapist (n=1), physical therapists (n=3) and a researcher educated within diagnostic radiography (n=1). All Panel 2 experts were from Europe, and the majority was women (n=6).

Vignette ranking surveys (individual survey and consensus survey)

All members of Panel 1 ranked the 25 patient vignettes according to the likelihood of the symptoms being due to hand OA. The distribution of rankings demonstrated a relative lack of agreement for most patients (figure 3). For example, the ranks for case 10 (a 63-year-old woman) showed an almost complete range from 2nd to 24th. The description of the woman is provided in online supplemental file S1. The agreement between experts was moderate (Kendall's coefficient of concordance=0.53).

In consensus, 17 experts agreed on a ranking of patients according to the likelihood of their symptoms being due to hand OA. The consensus ranking was strongly correlated with the mean ranking based on individual surveys (Spearman correlation coefficient=0.81). Key positive features that were deemed relevant in the expert panel discussion when ranking patients included: involvement of interphalangeal joints and thumb base joints, long symptom duration, the presence and severity of radiographic features, concordance between symptoms and radiographic or clinical OA features and higher age. Key negative features included: involvement of metacarpophalangeal (MCP) and wrist joints, high levels of
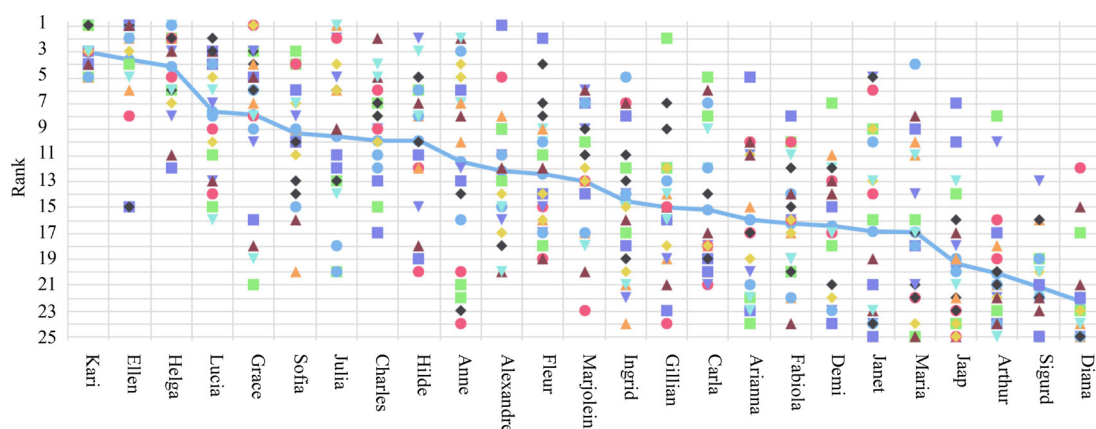

Figure 3 Ranking of the 25 patient vignettes from 1st (most likely) to 25th (least likely) based on the likelihood of their symptoms being due to hand OA. Each square or circle represents the rank of one expert. The blue line shows the mean rank across all experts. OA, osteoarthritis.

inflammatory biomarkers, skin psoriasis and prolonged morning stiffness.

Identification of target population, target joints, domains and categories

Identifying the population to which the classification criteria sets should be applied

To ensure that the criteria are being applied in a population with symptomatic (in contrast to asymptomatic) hand OA, the expert panel members agreed on two mandatory criteria: (1) the target population should include people with pain, aching or stiffness in at least one target joint and (2) their complaints should not be better explained by another disease or acute injury, since painful or stiff hand joints generally showed limited ability to discriminate between patients with and without hand OA in Phase 1.[2] Differential diagnoses vary among patients with different presentations. Nonetheless, they may include crystal arthropathies, non-inflammatory hand conditions such as haemochromatosis and systemic inflammatory joint diseases such as rheumatoid arthritis (RA) and psoriatic arthritis (PsA). Because PsA is especially challenging to distinguish from hand OA, the experts recommended that people with psoriasis should be excluded from the target population.

Symptom duration was initially one of the criteria in the 1000minds surveys (see the 'Domains of importance for the criteria sets' section). After the 1000minds preferences survey, the Panel 1 experts agreed that chronic symptoms should be mandatory. The first mandatory criterion was thus changed to 'pain, aching or stiffness in at least one target joint on most days in the previous 6 weeks'. Furthermore, patients with long morning stiffness of 60 min or longer were excluded from the target population (ie, second mandatory criterion) after the 1000minds validity exercises due to concerns that these patients had a systemic inflammatory joint disease.

Identifying target joints

The target joints differ depending on the criteria set being applied. For the overall hand OA criteria set, the target joints were defined as the bilateral second–fifth distal interphalangeal (DIP), second–fifth proximal interphalangeal (PIP), first interphalangeal (IP1) and thumb base joints. The second–fifth DIP, second–fifth PIP and IP1 joints were defined as target joints for the interphalangeal OA criteria set, whereas the thumb base joints were the target joints for the thumb base OA criteria set. In Phase 1, the ability to discriminate between patients with and without hand OA was similar for radiographic features in the DIP and PIP joints,[2] and the experts agreed that the DIP and PIP joints could be treated as one entity. For radiographic features in the thumb base joints, the experts agreed that the first carpometacarpal (CMC1) joints should be evaluated, and not the scaphotrapeziotrapezoidal (STT) joints, due to poorer discrimination for OA features in the STT joint between patients

with hand OA and controls and frequent co-occurrence of OA in the CMC1 and STT joints in Phase 1.[2]

Domains of importance for the criteria sets

The experts agreed on a list of criteria and their categories to be tested in the 1000minds vignette categorisation survey, including age, symptom duration, duration of morning stiffness, number of joints with osteophytes, number of joints with joint space narrowing, number of joints with symptomatic OA (eg, symptoms and radiographic features in the same joints) and inflammatory biomarkers (online supplemental file S2). Due to few other criteria describing the symptom characteristics of hand OA, 'stiffness after rest/inactivity' was added to the list of criteria after the consensus ranking exercise. Information about this symptom was added to the 25 patient vignettes by IKH using clinical judgement because information about this symptom was not collected in Phase 1.[2]

For the definition of short versus long morning stiffness, the highest AUC value was found when long morning stiffness in the finger joints was defined as more than 30 min (0.62). The cut-off value for short versus long inactivity stiffness was based on clinical experience due to lack of data to support the decision. For the radiographic and symptomatic hand OA criteria, we observed a higher proportion of patients with hand OA with higher joint counts. The chosen categories included a range of joints with a similar probability of having hand OA. Both erythrocyte sedimentation rate (ESR) and C reactive protein (CRP) were included as inflammatory biomarkers due to similar discriminatory ability in Phase 1.[2] The values were dichotomised (ESR: ≥ 15 mm/hour and CRP: ≥ 5 mg/L) based on the highest observed AUC (0.61 and 0.59, respectively).

Typical RA features, such as symptoms or swelling of MCP and wrist joints, were excluded from the list of key features. The experts agreed that wrist symptoms might be difficult to distinguish from thumb base symptoms, and OA may also lead to symptoms in MCP joints. Furthermore, the assessment of swelling would require a clinical examination. Symptoms or swelling in MCP and wrist joints showed similar discriminatory capability as the inflammatory biomarkers.[2] Importantly, the absence of another inflammatory rheumatic and musculoskeletal disease does not mean that the person has OA. It was considered important to focus mainly on criteria typical for hand OA.

Vignette categorisation survey (individual surveys)

The 1000minds vignette categorisation survey was completed by 16 Panel 1 experts. They demonstrated good agreement except for the 'Symptomatic OA in target joints' criterion. The experts concluded that the criterion was complicated due to counting and combining both different radiographic features and symptoms (online supplemental file S2). Hence, the criterion was changed to a dichotomous item asking whether the person has radiographic OA in most symptomatic joints.

The criteria were further refined with a more explicit description of target joints and shorter descriptions of categories. The categories for all criteria were listed in the order from least likely to most likely associated with hand OA (table 1).

Preferences survey (individual surveys)

All Panel 1 experts (n=21) completed the 1000minds preferences survey for overall hand OA. Table 1 shows the initial weights for each of the eight criteria and their categories (weight 1). Radiographic features were considered most important. Despite receiving a relatively low weight in the preferences survey, the experts agreed that symptom duration of at least 6 weeks should instead be a mandatory criterion, as people with short-lived symptoms are not preferable in most hand OA clinical trials. The experts also suggested removing 'inactivity stiffness' and 'inflammatory biomarkers' due to potential overlap with the 'morning stiffness' criterion and concerns about its feasibility, respectively. Although the trade-off questions asked by the PAPRIKA method depend on the criteria that are included, on average, 90% of the questions to be answered by each expert were the same after removing these three criteria. It was, therefore, not necessary to re-run the survey. After the three criteria were removed, the weights of the remaining five criteria were automatically calculated and proportionately increased accordingly (table 1, weight 2). The experts were consulted again to confirm the face validity of the final criteria and weights.

The preferences surveys, based on five criteria, were repeated for interphalangeal OA and thumb base OA. All Panel 1 experts (n=21) completed both surveys. Their weights were similar to the weights from the survey for overall hand OA (online supplemental file S3).

Evaluation of validity

The Spearman correlation coefficient was 0.65, indicating good correlation between the consensus ranking from the 1000minds vignette ranking survey and the ranking based on the weights of the eight criteria that were tested in the 1000minds preferences survey. After removal of three criteria ('duration of symptoms', 'inactivity stiffness' and 'inflammatory biomarkers'), the correlation with the consensus ranking remained more or less unchanged (0.63), and the impact of these three removed criteria was considered small. The concordance between the two ranks is demonstrated in online supplemental file S4.

All Panel 1 experts (n=21) participated in the second validity exercise. The number of experts who wanted to include the patients in Set 2A in the hypothetical trial increased with higher total scores (figure 4). The Spearman correlation coefficient between the total score and the number of experts who wanted to include the patient in the hypothetical trial was very high (0.91). The experts repeated the validity survey for the interphalangeal OA and thumb base OA criteria sets (online

Table 1 Criteria for overall hand OA with the mean weight assigned to each criterion (bolded numbers) and the categories within each criterion

Criteria and categories	Mean weight 1*	Mean weight 2*
Age		
Below 45 years	0	0
45–54 years	5.9	8.1
55–64 years	9.5	13.0
65 years and above	**13.3**	**18.3**
Duration of symptoms in DIPs, PIPs, IP1 and thumb base joints		
Less than 6 weeks	0	–
6 weeks or longer	**7.7**	
Duration or morning stiffness in DIPs, PIPs, IP1 and thumb base joints		
Long (more than 30 min)	0	0
None	5.6	7.7
Short (30 min or less)	**10.7**	**14.7**
Duration of stiffness after inactivity in DIPs, PIPs, IP1 and thumb base joints		
Long (more than 5 min)	0	–
None	2.5	
Short (5 min or less)	**7.4**	
Number of DIPs, PIPs, IP1 and CMC1 joints with osteophytes		
None	0	0
1–2 joint(s)	9.0	12.4
3–5 joints	15.9	21.8
6 or more joints	**20.4**	**28.0**
Number of DIPs, PIPs, IP1 and CMC1 joints with joint space narrowing		
None	0	0
1–2 joint(s)	6.8	9.3
3–5 joints	11.0	15.1
6 or more joints	**14.8**	**20.3**
Radiographic OA (osteophytes or joint space narrowing) in at least 50% of the joints (DIPs, PIPs, IP1 and CMC1), in which the person has experienced pain, aching or stiffness on most days in the previous 6 weeks		
No	0	0
Yes	**13.6**	**18.7**
Inflammatory biomarkers		
Elevated CRP and ESR	0	–
Elevated CRP or ESR	3.4	
Normal CRP and ESR	**12.2**	

*Weight 1: initial weights for the 8 criteria tested in the 1000minds survey.
†Weight 2: weights for the 5 criteria after removal of 3 criteria. The sum of the weights for each criterion's most extreme category (bolded weights) across the criteria gives a total score of 100.
CMC1, first carpometacarpal; CRP, C reactive protein; DIP, distal interphalangeal; ESR, erythrocyte sedimentation rate; IP1, first interphalangeal; OA, osteoarthritis; PIP, proximal interphalangeal.
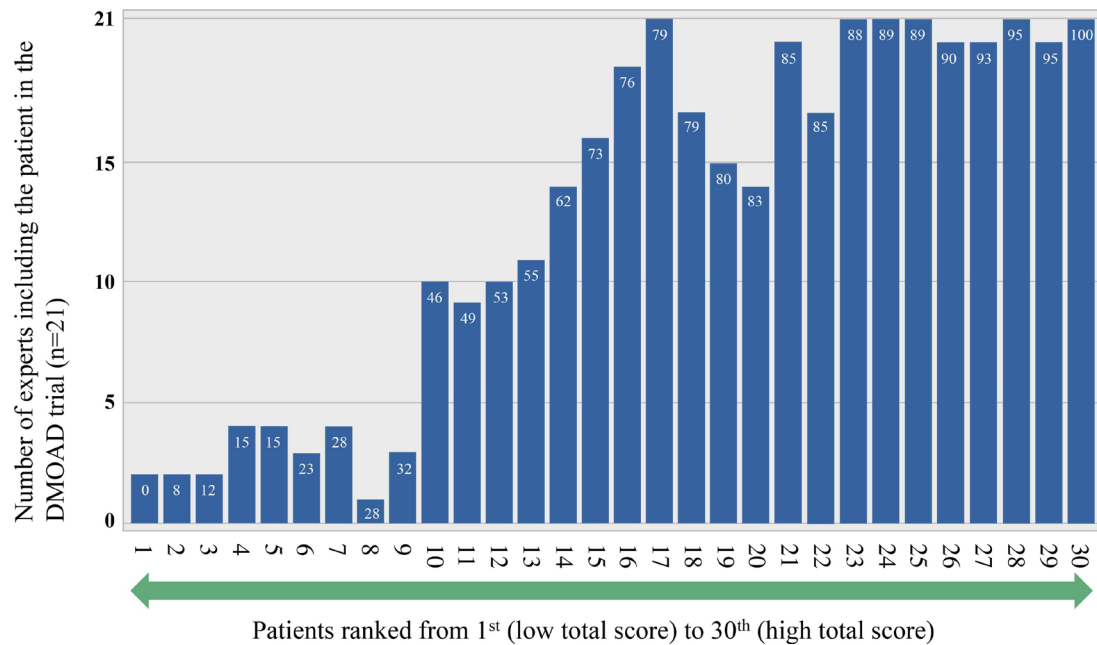
Figure 4 Proportions of experts who would enrol the patient in a hypothetical clinical trial of a possible disease-modifying drug for patient vignettes arranged from the lowest to the highest probability of being classified as having hand OA based on the calculated total score from the Phase 2 criteria set. The white numbers in the blue bars represent the total scores (range: 0–100) based on the criteria set presented in table 1 (weight 2). DMOAD, disease-modifying osteoarthritis drug; OA, osteoarthritis.

supplemental file S5). The Spearman correlation coefficients were 0.80 and 0.82 for interphalangeal OA and thumb base OA, respectively. It was noted that experts tended to not include patients with long morning stiffness (eg, 1 hour or longer) in the hypothetical clinical trial. Hence, the experts agreed to exclude patients with long morning stiffness of 1 hour or longer from the target population. The category for 'long morning stiffness' was changed to '31–59 min' accordingly.

Evaluation of reliability

The reliability of the five proposed criteria and their categories, as shown in table 1, was excellent between the Panel 2 experts (n=7). The multirater kappa value ranged from 0.90 for radiographic OA in symptomatic joints to 0.99 for osteophytes and age. Similarly, the percentage agreement ranged from 94% to 99%. The ICC value was 1.00.

DISCUSSION

To develop new hand OA classification criteria, we used a consensus-based decision analysis approach to identify the criteria and their categories, and determine their weights for differentiating between people with and without hand OA.

Initially, we asked our experts to rank patient vignettes according to the likelihood of the symptoms being due to hand OA. Our results emphasise that experts have very different opinions about which features are the most relevant when determining whether the symptoms are due to hand OA or not. Hence, it is clear that we need

a better tool for disease classification when determining the relative importance of key features.

Using 1000minds software, we identified five criteria that are important when determining whether symptoms in a person are due to hand OA or not. In addition, two criteria were considered mandatory: (1) pain, aching or stiffness in at least one hand joint on most days in the previous 6 weeks and (2) the absence of a disease or acute trauma that could better explain the symptoms. We expect the criteria set to be frequently used in clinical trials, and having symptomatic OA was deemed essential in these settings. Although most patients with and without hand OA in Phase 1 had experienced symptoms for more than 6 months,[2] the criterion about the duration of symptoms was dichotomised using 6 weeks or longer as a cut-off for chronic complaints. The same cut-off was also used in the American College of Rheumatology (ACR) and EULAR classification criteria set for RA.[9] The long symptom duration among participants in Phase 1 is likely explained by recruitment of patients from secondary/tertiary care. A cut-off at 6 weeks will likely be more applicable in primary care settings.

Fulfilment of the second mandatory criterion about no other diseases or acute injuries explaining the symptoms requires clinical expertise. The experts did not provide a comprehensive list of relevant differential diagnoses or tests that should be performed to exclude them, as this would have been beyond the scope of the classification criteria. The physician should be responsible for the evaluation of differential diagnoses, which will vary across different populations. In a clinical trial, fulfilling the

second mandatory criterion is crucial because important differential diagnoses such as RA and PsA have different pathogenesis requiring another treatment. We acknowledge that in large observational cohort studies, an expert evaluation of all participants may be challenging. The fulfilment of the item may be based on self-reported data on important differential diagnoses, which the investigators can determine based on the study population.

The remaining five criteria were age, morning stiffness, radiographic osteophytes, radiographic joint space narrowing and an evaluation of whether symptoms and radiographic features occurred in the same hand joints. Weights for these criteria and their categories, representing their relative importance, were determined using the 1000minds software. When we applied the resulting scoring system to our patient vignettes, the ranking was close to the consensus ranking produced by our experts based on their clinical experience and judgement. The preliminary scoring system was also in line with Phase 1 results, supporting its face and construct validity. The experts agreed that the same criteria were important across all criteria sets, and did not repeat all 1000minds surveys to develop the criteria sets for interphalangeal OA and thumb base OA. In line with this decision, the 1000minds preferences survey revealed similar weights for the five criteria across the three proposed criteria sets.

The absence of long morning stiffness of 1 hour or above was listed as a mandatory criterion due to concerns that patients with long morning stiffness are likely to have systemic inflammatory joint disease. Morning stiffness in OA is generally considered to be of short duration. Among patients with hand OA in Phase 1, most patients with morning stiffness had a duration of 30 min or less.[2] In line with these results, the ACR criteria set for knee OA gives one point to people with morning stiffness less than 30 min.[10]

Validation of the preliminary sets of criteria was performed using data from Phase 1, which was used to identify important features, and experts who were familiar with the process. This process may have biased the results, and validation of the final criteria set in an independent sample is crucial.

In Phase 2, we used the 1000minds software to integrate decision analysis, which is considered more transparent and flexible than Delphi consensus approaches. Nonetheless, the results will depend on the expertise of the expert panels. Our experts were carefully selected to provide a broad range of international and multidisciplinary expertise to ensure different viewpoints and perspectives in the discussions and the surveys.

To conclude, the work in Phase 2 produced three preliminary sets of criteria for overall hand OA, interphalangeal OA and thumb base OA. The associated weights of the criteria and their categories, representing their relative importance, were developed using decision-making software (1000minds) and was informed by Phase 1 results, if needed. These results will inform the three final criteria sets. In the final phase of the process of developing new classification criteria for hand OA, a cut-off to be used to define hand OA and the subsets will be determined together with a preliminary validation of the criteria sets.

Author affiliations
[1]Division of Rheumatology and Research, Diakonhjemmet Hospital, Oslo, Norway
[2]Rheumatology section, Boston University School of Medicine, Boston, Massachusetts, USA
[3]Arthritis Research UK Epidemiology Unit, National Institute for Health Research Biomedical Research Centre, The University of Manchester, Manchester, UK
[4]Academic Rheumatology, School of Medicine, University of Nottingham, Nottingham, UK
[5]NIHR Nottingham Biomedical Research Centre, Nottingham, UK
[6]INSERM CRSA, Sorbonne University, Paris, France
[7]Department of Rheumatology, Hopital Saint-Antoine, Paris, France
[8]Primary Care Centre Versus Arthritis, School of Medicine, Keele University, Keele, UK
[9]Department of Rheumatology, Instituto de Investigación Sanitaria Fundación Jimenez Díaz, Universidad Autonoma de Madrid, Madrid, Spain
[10]Rheumatology Department, Queen Elizabeth Hospital, Woodville, South Australia, Australia
[11]Faculty of Health and Medical Sciences, Adelaide Medical School, The University of Adelaide, Adelaide, South Australia, Australia
[12]Department of Medicine, Division of Rheumatology, Cedars-Sinai Medical Center, Los Angeles, California, USA
[13]Department of Rheumatology, Landspitali, Reykjavik, Iceland
[14]Department of Surgery, Central Finland Central Hospital, Jyväskylä, Finland
[15]Department of Rheumatology and Immunology, Singapore General Hospital, Singapore
[16]National Advisory Unit on Rehabilitation in Rheumatology, Division of Rheumatology and Research, Diakonhjemmet Hospital, Oslo, Norway
[17]Rheumatology Unit, Department of Medicine, University of Padua, Padova, Italy
[18]Section for Outcomes Research, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Wien, Austria
[19]Institute for Arthritis and Rehabilitation, Ludwig Boltzmann Gesellschaft, Wien, Austria
[20]Division of Rheumatology, University of Debrecen, Debrecen, Hungary
[21]Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands
[22]Department of Plastic, Reconstructive and Hand Surgery, Amsterdam UMC, Amsterdam, The Netherlands
[23]Department of Rheumatology, Ghent University, Ghent, Belgium
[24]Clinical Epidemiology Unit, Orthopedics, Department of Clinical Sciences Lund, Lund University, Lund, Sweden
[25]Department of General Practice, Erasmus University Rotterdam, Rotterdam, The Netherlands
[26]Department of Economics, University of Otago, Dunedin, New Zealand
[27]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

ORCID iDs
Ida K Haugen http://orcid.org/0000-0001-7810-2216
David Felson http://orcid.org/0000-0002-2668-2447
Francis Berenbaum http://orcid.org/0000-0001-8252-7815
Rikke Helene Moe http://orcid.org/0000-0001-7601-5346
Roberta Ramonda http://orcid.org/0000-0002-9683-8873
Valentin Ritschl http://orcid.org/0000-0001-8763-8215
Tanja A Stamm http://orcid.org/0000-0003-3073-7284
Zoltan Szekanecz http://orcid.org/0000-0002-7794-6844
Ingvild Kjeken http://orcid.org/0000-0002-3971-2852
Nina Osteras http://orcid.org/0000-0001-8602-342X
M Marshall http://orcid.org/0000-0001-8163-6948

REFERENCES

1 Kloppenburg M, Kwok W-Y. Hand osteoarthritis--a heterogeneous disorder. *Nat Rev Rheumatol* 2011;8:22–31.
2 Haugen IK, Felson DT, Abhishek A, *et al*. Development of classification criteria for hand osteoarthritis: comparative analyses of persons with and without hand osteoarthritis. *RMD Open* 2020;6:e001265.
3 Neogi T, Aletaha D, Silman AJ, *et al*. The 2010 American College of Rheumatology/European League against rheumatism classification criteria for rheumatoid arthritis: phase 2 methodological report. *Arthritis Rheum* 2010;62:2582–91.
4 Rider LG, Ruperto N, Pistorio A, *et al*. 2016 ACR-EULAR adult dermatomyositis and polymyositis and juvenile dermatomyositis response criteria-methodological aspects. *Rheumatology* 2017;56:1884–93.
5 Tedeschi SK, Johnson SR, Boumpas DT, *et al*. Multicriteria decision analysis process to develop new classification criteria for systemic lupus erythematosus. *Ann Rheum Dis* 2019;78:634–40.
6 Johnson SR, Naden RP, Fransen J, *et al*. Multicriteria decision analysis methods with 1000Minds for developing systemic sclerosis classification criteria. *J Clin Epidemiol* 2014;67:706–14.
7 Taylor WJ, Fransen J, Jansen TL, *et al*. Study for updated gout classification criteria: identification of features to classify gout. *Arthritis Care Res* 2015;67:1304–15.
8 Hansen P, Ombler F. A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. *J Multi-Crit Decis Anal* 2008;15:87–107.
9 Aletaha D, Neogi T, Silman AJ, *et al*. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League against rheumatism collaborative initiative. *Ann Rheum Dis* 2010;69:1580–8.
10 Altman R, Asch E, Bloch D, *et al*. Development of criteria for the classification and reporting of osteoarthritis: classification of osteoarthritis of the knee. *Arthritis & Rheumatism* 1986;29:1039–49.