Title: Discrete Choice Experiment on a Magnetic Resonance Imaging Scoring System for Temporomandibular Joints in Juvenile Idiopathic Arthritis

Running Head: Discrete Choice Experiment on JAMRIS-TMJ

Authors:

Mirkamal Tolend, BSc, The Hospital for Sick Children and University of Toronto, Toronto, ON, Canada

Thitiporn Junhasavasdikul, MD, Ramathibodi Hospital, Bangkok, Thailand

Randy Q. Cron, MD, PhD, Children's of Alabama, Birmingham, AL, United States

Emilio J. Inarejos Clemente, MD, Hospital Sant Joan de Deu, Barcelona, Spain

Thekla von Kalle, MD, Olgahospital Klinikum Stuttgart, Stuttgart, Germany

Christian J. Kellenberger, MD, University Children's Hospital Zürich, Zürich, Switzerland

Bernd Koos, DMD, University Hospital Tübingen, Tübingen, Germany.

Elka Miller, MD, CHEO, University of Ottawa, ON, Canada

Marion A. van Rossum, MD, PhD, Emma Children's Hospital, Academic Medical Centre, and Amsterdam Rheumatology and Immunology Center, Reade, Amsterdam, The Netherlands

Rotraud K. Saurenmann, MD, University Children's Hospital Zürich, Zürich, Switzerland

Lynn Spiegel, MD, The Hospital for Sick Children, Toronto, ON, Canada.

Jennifer Stimec, MD, The Hospital for Sick Children, Toronto, ON, Canada.

Marinka Twilt, MD, MScE, PhD, Alberta Children's Hospital, and University of Calgary, Calgary, Alberta, Canada

Nikolay Tzaribachev, MD, PhD, Pediatric Rheumatology Research Institute, Bad Bramstedt, Germany

Shelly Abramowicz, DMD, MPH, Emory University School of Medicine, and Children's Healthcare of Atlanta, Atlanta, GA, United States

Simone Appenzeller, MD, PhD, University of Campinas, Campinas, Brazil

Linda Z. Arvidsson, DDS, PhD, University of Oslo, Oslo, Norway

Saurabh Guleria, MD, MD, Austin Radiological Association, Austin, TX, United States

Jacob L. Jaremko, MD, PhD, University of Alberta, Edmonton, AB, Canada

Eva Kirkhus, MD, PhD, Oslo University Hospital, Oslo, Norway

Tore A. Larheim, DDS, PhD, University of Oslo, Oslo, Norway

Arthur B. Meyers, MD, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

Jyoti Panwar, MD, Christian Medical College and Hospital, Vellore, Tamil Nadu, India

Cory M. Resnick, DMD, MD, Boston Children's Hospital, Boston, MA, United States

Susan C. Shelmerdine, MBBS, Great Ormond Street Hospital, London, UK

Brian M. Feldman, MD, MSc, The Hospital for Sick Children, Toronto, ON, Canada

*Andrea S. Doria MD, PhD, MSc, MBA (corresponding author), The Hospital for Sick Children, Toronto, ON, Canada

Corresponding author contact:

Andrea S. Doria, MD, PhD, MSc, MBA
Professor, Associate Vice-Chair of Research (Injury, Repair and Inflammation), Department of Medical Imaging, University of Toronto
Radiologist, Senior Scientist, Research Director, Department of Diagnostic Imaging
The Hospital for Sick Children
555 University Avenue, 2nd floor
Toronto, ON M5G1X8
Phone: 416-813-6079
Fax: 416-813-7591
Email: Andrea.Doria@sickkids.ca

Conflict of Interest: The authors do not have any conflict of interest related to this work.

Financial Support: This study did not receive any specific financial support from public or commercial sources.

Word Count: 3093

Abbreviations:

DCE – Discrete Choice Experiment; ICC – Intraclass Correlation Coefficient; JAMRIS-TMJ – Juvenile Idiopathic Arthritis Magnetic Resonance Imaging Scoring System for Temporomandibular Joints; JIA – Juvenile Idiopathic Arthritis; MRI – Magnetic Resonance Imaging; TMJ – Temporomandibular Joints.

Abstract:

Objective. To determine the relative importance weights of items and grades of a newly developed additive outcome measure called the juvenile idiopathic arthritis (JIA) magnetic resonance imaging (MRI) scoring system for temporomandibular joints (TMJ, JAMRIS-TMJ).

Methods. An adaptive partial-profile discrete choice experiment (DCE) survey using the 1000Minds platform was independently completed by members of an expert group consisting of radiologists and non-radiologist clinicians to determine the group-averaged relative weights for JAMRIS-TMJ. Subsequently, an image-based vignette ranking exercise was done, during which experts individually rank-ordered 14 patient vignettes for disease severity while blinded to the weights and unrestricted to JAMRIS-TMJ assessment criteria. Validity of the weighted JAMRIS-TMJ was tested by comparing the consensus-graded, DCE-weighted JAMRIS-TMJ score of the vignettes with their unrestricted image-based ranks provided by the experts.

Results. Nineteen experts completed the DCE survey and 21 completed the vignette ranking exercise. Synovial thickening and joint enhancement showed higher weights per raw score compared to bone marrow items and effusion in the inflammatory domain, while erosions and condylar flattening showed non-linear and higher weights compared to disk abnormalities in the damage domain. The weighted JAMRIS-TMJ score of the vignettes correlated highly with the ranks from the unrestricted comparison method, with median Spearman's rho of 0.92 (intra-quartile range: 0.87-0.95) for the inflammation and 0.93 (0.90-0.94) for the damage domain.

Conclusions. A DCE survey was used to quantify the importance weights of the items and grades of the JAMRIS-TMJ. The weighted score showed high convergent validity with an unrestricted, holistic vignette ranking method.

Significance and Innovation:

- A discrete choice experiment was used to develop a weighting scheme for the items and grades of a newly developed MRI scoring system for assessing the inflammation and damage in the temporomandibular joints of children with juvenile idiopathic arthritis (JAMRIS-TMJ).
- In the inflammatory domain of the scoring system, the importance weights for joint enhancement (34% of domain score) and synovial thickening (31%) were higher than the bone marrow items (9% and 10%) and effusion (16%).
- In the damage domain, erosions and condylar flattening were both weighted higher compared to disk abnormalities (38% and 49% vs 13%).
- The weighted JAMRIS-TMJ score showed high convergent validity when compared to an unrestricted image-based method of ranking vignettes (median Spearman's rho of 0.92 and 0.93 for the two domains).

Keywords: Magnetic Resonance Imaging, Juvenile Idiopathic Arthritis, Discrete Choice Experiment, Outcome Measure, Temporomandibular Joints

Introduction

Juvenile idiopathic arthritis (JIA) is the most common form of chronic arthritis in children and youth, with a prevalence of 1 in 1,000 children worldwide (1). In large consecutive series of JIA patients, approximately 40% have been found to develop some degree of inflammation and structural changes in the temporomandibular joint (TMJ) (2–4). While TMJ arthritis can be asymptomatic (5), it was recently reported that orofacial pain and functional disability are common and seem to persist over time in most patients, negatively impacting oral health-related quality of life (6). Early detection of TMJ arthritis may facilitate intervention to prevent joint damage and dysfunction.

TMJ arthritis cannot be assessed comprehensively by physical examination, ultrasound, conventional radiographs, or computed tomography imaging (7–13). Contrast-enhanced magnetic resonance imaging (MRI) remains the best available diagnostic tool as it allows for visualization of both soft tissue and osteochondral changes in the TMJ. Since many early changes are subtle, the evaluation of TMJ MRI remains subjective and necessitates a standardized and feasible outcome measure. To this end, the Juvenile Idiopathic Arthritis MRI (JAMRI) working group within the Outcome Measures in Rheumatology (OMERACT) research network has recently developed the JAMRI scoring system for TMJ (JAMRIS-TMJ) (14).

The JAMRIS-TMJ is constructed as a multi-item, additive outcome measure with each joint graded by inflammatory and damage domains. Once the scoring items and feasible grading criteria are defined, the relative importance weights of the items and their grades must be determined and validated for deriving composite domain scores. For example, studies have identified that mild levels of effusion and synovial enhancement are not specific to TMJ arthritis (15–17), emphasizing that MRI-observable features and their levels have different and context-

specific importance when interpreting the MRI of TMJs. A discrete choice experiment (DCE) is helpful in this regard, offering a formalized and quantitative approach for eliciting the opinions of an expert panel in defining the relative importance weights of items in this type of measure (18–21). For a brief background on DCE, refer to Supplementary file 1.

In this study, we determined the relative importance weights of item and grades of the JAMRIS-TMJ using a DCE survey (22). The resulting weighting scheme enables the calculation of percentagewise inflammation and damage domain scores using the JAMRIS-TMJ method of MRI evaluation. To test the validity of the elicited weights, we conducted a vignette ranking exercise. The weighted JAMRIS-TMJ score ranking approach was tested against a holistic, image-to-image comparison approach as the reference standard, since the latter method allows greater differentiation and does not entail the reductionistic assumptions of the DCE process nor the restrictions inherent in the JAMRIS-TMJ grading criteria. The specific aims of the study were: (1) to determine the relative importance weights of the items and grades in the JAMRIS-TMJ using an adaptive DCE method within a multicenter, multi-specialty group of experts; (2) to assess the validity of the DCE-derived importance weights using an image vignette-based exercise, by testing the correlation of the JAMRIS-TMJ weighted vignette score with the vignette rank given through a scoring system-independent method.

Methods

This study was approved by the Research Ethics Board (REB) of The Hospital for Sick Children (Toronto, Canada, study reference 1000042164). Information letters were provided to the participants before each activity to explain the study and that their voluntary completion and

submission of the study surveys constituted their implied consent to participate in the study. Considering the practical limitations, and that the imaging exams used for creating the vignettes were anonymized and retrospective in nature, written consent requirement was waived by the REB. The study was conducted in two phases, the first being the DCE survey to develop the relative importance weights, and the second the vignette ranking exercise which tested the face and convergent validity (23) of the DCE weighted scoring system. Figure 1 summarizes the methods in a flowchart.

Discrete Choice Experiment Survey

An adaptive, partial-profile DCE survey administered through the 1000Minds software (22) was completed independently and anonymously by a multidisciplinary group of experts. Radiologists and other clinicians were invited if they routinely assess TMJ MRI in JIA patients. Each expert participant completed separate DCE surveys for the inflammatory and damage domains. All discrete choice questions asked the expert to compare two hypothetical sets of findings with different, non-dominating grades in the same two JAMRIS-TMJ domain items, and choose which scenario represented “more severe disease, assuming all else being equal”, or rate them as equal (Supplementary file 2). The relative weights were derived by the 1000Minds software utilizing the PAPRIKA (Potentially All Pairwise Rankings of all possible Alternatives) method (22). A complete set of item and grade importance weights was obtained for each DCE survey participant. The individual sets of weights were averaged over the entire group of experts to serve as the relative weights for the scoring system for testing. The weights were kept hidden until after the ranking exercise.

Vignette Ranking Exercise

Convergent validity of the weighted JAMRIS-TMJ was tested through a vignette ranking exercise conducted by a multidisciplinary group of radiologists and other non-radiologist clinicians within the JAMRI research network. Fourteen vignettes representing single TMJs from JIA patients were constructed from representative slices from each of the six imaging sequences from a TMJ MRI protocol for JIA utilizing dedicated surface coils (Supplementary file 4). The six images consisted of three pre-contrast sequences (fat suppressed sagittal oblique T2, sagittal oblique proton density-weighted, and coronal T1-weighted) and three gadolinium-enhanced T1-weighted fat suppressed sequences in three planes (axial, sagittal oblique, and coronal). Participants independently ranked these vignettes in increasing order of severity of inflammation and osteochondral damage, allowing for tied ranks. Item-wise grades of the 14 vignettes achieved by consensus of two radiologists were provided for a subgroup of clinician participants who do not regularly interpret TMJ MRI exams themselves, hence they ranked graded images. To simulate a pragmatic and holistic method of vignette-to-vignette comparison that is independent of any scoring method, all participants were instructed not to base their ranking on any summation of scores, allowing for the possibility that more important items or certain combinations of item-grades can disproportionally influence the disease severity ranking.

The item-wise JAMRIS-TMJ raw scores for each of the 14 vignettes were decided by consensus during a face-to-face and video conference meeting among a subgroup of participants (n=11) who regularly interpret TMJ MRI examinations. Weighted JAMRIS-TMJ scores for the 14 vignettes were produced using these consensus grades and the importance weights derived from the DCE. The weighted JAMRIS-TMJ score was then correlated with the ranking provided by each of the participating experts. This correlation tested the combined impact of several factors

related to the face and content validity of weighted JAMRIS-TMJ, including the items, grades, and their relative weights; the joint factor independence (21), transitivity and other assumptions of the adaptive partial-profile DCE method used to derive the weights (22); as well as the discriminative capacity and feasibility of the grading criteria.

Sample Size Considerations

The adaptive DCE method from 1000Minds which we used for generating the weights provides a complete set of weights for each item and grade level of the scoring system for every participant (for details, see supplementary file 1 and 3) (22). Therefore, the sample size requirement for the number of participants was not based on quantitative simulations for model convergence, but instead, on achieving a comprehensive and saturated opinion base that is representative of the level of heterogeneity among clinicians from multiple centers and specialties. Convenience sampling from an international research interest group was used to enroll experts from multiple specialties for the two study exercises. The number of vignettes used for the ranking was also subjectively determined to provide a balance between representing the common item combinations across the spectrum of two disease domains and reducing respondent error.

Statistical Analysis

In the DCE survey, homogeneity of the relative weights within the expert group was assessed in two ways. First, the representativeness of the group-averaged set of relative weights was tested by calculating the Spearman's rank correlation coefficients (rho) between rankings of all potential item combinations produced by group-averaged weights and each of the participants' weights (22). Second, the agreement of the relative weights among the participants was assessed by calculating the intraclass correlation coefficients (ICC, two-way random, single measure,

absolute agreement type). In the vignette ranking exercise, agreement in the vignette rankings among the participants was assessed visually per vignette by scatterplots, and quantitatively by calculating the ICC of ranks given to each of the 14 vignettes. Spearman's rho was used for correlating the image-based ranking with the weighted JAMRIS-TMJ score. For both correlation coefficients, values ≤0.4 were defined as poor correlation, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and >0.81 as high correlation. Statistical analyses were performed using SPSS version 23 (IBM Corp., Armonk, NY). For further details regarding the DCE survey and the statistical tests used, refer to supplementary file 3.

Results

Nineteen experts completed the DCE survey in total, including 11 pediatric or maxillofacial radiologists, seven pediatric rheumatologists, six of whom self-identifying as not regularly interpreting TMJ MRI themselves, and one orthodontist, yielding 19 sets of item and grade weights. Approximately 20-25 discrete choice questions (Supplementary file 2) were required to obtain a full set of relative importance weights for the two domains of the scoring system for each participant (Figures 2 and 3): the five-item inflammatory domain required between 14-18 questions, and the three-item damage question required between 5-7. The number of questions varied between the participants due to the differences in opinion and the order in which the questions are presented (for more details, refer to supplementary files 1 and 3). Quantitative indices of the group's homogeneity on these weights were sufficiently high: the ranking of all possible combinations of items that is produced from each expert's weights correlated highly with the rank produced by the group-averaged weights, with median Spearman's rho=0.96 for the inflammation domain (interquartile range, or IQR:0.93-0.96), and 0.97 for the damage

domain (IQR:0.95-0.99); group-wide 19-rater agreement on these 8 and 5 non-zero weights for the two domains were substantial, at 0.71 for the inflammatory domain weights, and 0.77 for the damage domain. Therefore, the average of the 19 sets of weights from the experts were deemed representative to be used as the JAMRIS-TMJ weights (Table 1), which were kept hidden prior to the vignette ranking exercise.

Twenty-one experts consisting of 11 pediatric or maxillofacial radiologists, seven pediatric rheumatologists, two oral and maxillofacial surgeons, and one orthodontist (13 overlapping with the experts who participated in the DCE) completed the vignette ranking exercise. Overall, the ranks given to the 14 vignettes correlated substantially among the 21 participants, with the ICC of the inflammatory domain vignette ranking at 0.85, and the damage domain ranking at 0.91.

The group-averaged relative weights from the DCE survey revealed several differences between the items of the JAMRIS-TMJ and their grade levels (Table 1). Highest grade joint enhancement showed a 34% relative weight for assessment of inflammation, compared to the highest-grade joint effusion (16%) and bone marrow enhancement (10%). Condylar flattening and erosions showed non-linear changes between grade levels, with the second grade-level being weighted higher per score than the first. Differences between the radiologists and non-radiologist clinicians on the relative weights were not statistically different when adjusted for multiple testing. The participants agreed that the group-averaged set of importance weights seem to be an appropriate representation of the group's opinion for use in subsequent construct validity studies, and justifiable considering the current understanding of TMJ arthritis, and the sensitivity and specificity of observable items in contrast enhanced MRI. Nevertheless, in examining the range of potential item combinations for the two domains – up to 108 and 18 for the inflammatory and damage domain, respectively – it was identified that three out of four potential item-grade

combinations between weighted scores of 52% to 78% may be quite rare or impossible to obtain: grade 2 flattening and grade 1 erosions (59%), grade 1 flattening and grade 2 erosions (66%), both with no disk abnormalities, and grade 2 erosions with disk abnormalities but no flattening (62%).

The consensus DCE weighted JAMRIS-TMJ score for the 14 vignettes correlated very highly with the 21 sets of vignette ranks generated from the image-based ranking exercise, with median Spearman's rho of 0.92 (IQR:0.87-0.95) for the inflammatory domain, and 0.93 (IQR:0.90-0.94) for the damage domain (Figure 4). Vignettes which received weighted scores placing midway in disease severity spectrum showed more variability in the image-based ranking than those with weighted scores near the two extremes (Supplementary file 4). No significant subgroup differences were observed between the participants who performed the image-only ranking (those who self-identified as reading TMJ MRI regularly, n=15) versus those who performed the graded image ranking (those who do not usually interpret TMJ MRI themselves, n=6).

The full-profile comparison of the patient vignettes was not restricted in terms of the items and grading cut-offs of the scoring system, allowing higher levels of differentiation between disease stages, and therefore a greater potential for disagreement in vignette ranks between the two methods. During the post-exercise discussions, it was identified that there were subtle but appreciable differences in the image-based ranking of the vignettes which were not differentiated by change in the JAMRIS-TMJ score. These scenarios could be described as "high" grade 1 vs "low" grade 1 within the confines of the grading threshold. In vignettes with unreliable, borderline grading (e.g., considering a feature as "high grade 1" or "low grade 2"), or joints in which not knowing the patient age or the inability to compare with the contralateral TMJ

challenged the interpretation of certain potentially patient-specific findings, such as condylar flattening and bone marrow changes.

Discussion

In this study, we used an adaptive partial-profile discrete choice experiment (DCE) method to formalize the assignment of quantitative importance weights to the items and grades of the juvenile idiopathic arthritis MRI scoring system for temporomandibular joints (JAMRIS-TMJ). Synovial thickening and joint enhancement items were considered by the expert panel on average twice as important per raw score compared to the other three inflammatory domain items (Figure 2 and Table 1). This finding underlines the diagnostic importance for contrast administration for assessment of TMJs, although it has become more restricted in clinical practice due to potential concerns with cumulative deposition of Gadolinium in the body (24,25). In the damage domain, erosions were weighted the most important, followed by condylar flattening, both with non-linear per score weights, then disk abnormalities (Figure 3 and Table 1). The non-linear increase in weights of grades for these damage domain items better represents the ordinal scaling of the grading definitions for these items compared to unweighted scoring. In general, the weighting scheme represents the features of both the progressive and additive TMJ MRI scoring systems that the JAMRIS-TMJ was derived from (26–28), emphasizing the diagnostic features with higher specificity for active inflammation, while still allowing for further differentiation by ancillary items.

The JAMRIS-TMJ grading method focuses on measuring the items as independently as possible. Synovial thickening is measured only on fluid sensitive sequences as presenting with

intermediate signal intensity on MRI, pockets of fluid need to be considered in grading joint enhancement to distinguish them from enhanced synovium, and the bone marrow edema signal is considered only on pre-contrast images (14). To the extent that the items can be measured independently, and that the various combinations of these items are realistic and informative, it should be useful to add these items to produce composite domain scores. For example, a region of synovium that does not enhance after contrast may suggest residual pannus from prior disease that is not currently inflamed, differentiating it from active disease. However, practical issues still can cause correlation or restriction of grades between items. When there is severe structural damage in the joint, some soft tissue components, such as inactive pannus, become difficult to identify and grade. Disease may also be overestimated when a given finding cannot be reliably attributed to a specific item: differentiation of soft tissue components is difficult if not impossible to assess using only post-contrast images, and comparing post-contrast with corresponding pre-contrast images may not always be helpful. Grading of these changes may be improved with the utilization of measurement aids for different stages of joint inflammation and degeneration such as by using an imaging atlas (29).

Non-linearity in the change in weighted JAMRIS-TMJ score between adjacently ranked vignettes (see Supplementary file 5) likely resulted from the limited vignette selection. However, it may also suggest that some theoretical combination of item scores in these intervals are too rare or transient to be captured. A cross-sectional study using the scoring system on a large consecutive series of patients would be helpful to study the true prevalence in these intervals of the scoring spectrum.

The chief limitation of this study was that the number of vignettes which could be rank-ordered was relatively low, precluding a more complex study design that could directly quantify the

advantage of weighted scores over unweighted scores when correlating to the holistic, image-based rank. To achieve this, it would be necessary to select the vignettes such that the difference between the raw score and the weighted scores is maximized, allowing for a more efficient differentiation between the two correlations. Increasing the number of vignettes to serve this purpose would also be challenging, since it would increase the cognitive burden of ranking, potentially leading the participants to use simplifying heuristics in their comparison of vignettes and hence skewing the way they applied the relative weights. Instead, the vignettes were selected to better represent the various common presentations across the entirety of the scoring spectrum in both domains, thus capturing more of the nuances in item combinations.

Conclusions

The DCE survey facilitated the development of relative importance weights of items and grades in the JAMRIS-TMJ, which showed high convergent validity with a holistic, scoring system-independent method of image assessment when applied to rank a series of TMJ MRI vignettes. The relative weights derived from the DCE revealed differences between the items as well as between the different grades of items, which would not be captured by the number of grades allotted to the items. The weighting scheme is therefore crucial for scaling the JAMRIS-TMJ inflammatory and damage domain scores in accordance with the perceived differences in the items and their grade levels, enabling their application as standardized outcome measures in clinical practice and research including clinical trials in JIA. Our methodology combining adaptive DCE with validation by subsequent holistic vignette ranking exercise could be applied to relative weighting of components of other imaging-based grading systems.

References

1. Manners PJ, Bower C. Worldwide prevalence of juvenile arthritis why does it vary so much? *J Rheumatol* 2002;29:1520–1530.

2. Larheim TA, Doria AS, Kirkhus E, Parra DA, Kellenberger CJ, Arvidsson LZ. TMJ imaging in JIA patients—An overview. *Semin Orthod* 2015;21:102–110.

3. Cannizzaro E, Schroeder S, Müller LM, Kellenberger CJ, Saurenmann RK. Temporomandibular Joint Involvement in Children with Juvenile Idiopathic Arthritis. *J Rheumatol* 2011;38:510–515.

4. Stoll ML, Sharpe T, Beukelman T, Good J, Young D, Cron RQ. Risk factors for temporomandibular joint arthritis in children with juvenile idiopathic arthritis. *J Rheumatol* 2012;39:1880–1887.

5. Twilt M, Mobers SMLM, Arends LR, Cate R ten, Suijlekom-Smit L van. Temporomandibular involvement in juvenile idiopathic arthritis. *J Rheumatol* 2004;31:1418–1422.

6. Rahimi H, Twilt M, Herlin T, Spiegel L, Pedersen TK, Küseler A, et al. Orofacial symptoms and oral health-related quality of life in juvenile idiopathic arthritis: a two-year prospective observational study. *Pediatr Rheumatol Online J* 2018;16:47.

7. Weiss PF, Arabshahi B, Johnson A, Bilaniuk LT, Zarnow D, Cahill AM, et al. High prevalence of temporomandibular joint arthritis at disease onset in children with juvenile idiopathic arthritis, as detected by magnetic resonance imaging but not by ultrasound. *Arthritis Rheum* 2008;58:1189–1196.

8. Munir S, Patil K, Miller E, Uleryk E, Twilt M, Spiegel L, et al. Juvenile idiopathic arthritis of the axial joints: a systematic review of the diagnostic accuracy and predictive value of conventional MRI. *AJR Am J Roentgenol* 2014;202:199–210.

9. Koos B, Twilt M, Kyank U, Fischer-Brandies H, Gassling V, Tzaribachev N. Reliability of clinical symptoms in diagnosing temporomandibular joint arthritis in juvenile idiopathic arthritis. *J Rheumatol* 2014;41:1871–1877.

10. Muller L, Kellenberger CJ, Cannizzaro E, Ettlin D, Schraner T, Bolt IB, et al. Early diagnosis of temporomandibular joint involvement in juvenile idiopathic arthritis: a pilot study comparing clinical examination and ultrasound to magnetic resonance imaging. *Rheumatol Oxf Engl* 2009;48:680–5.

11. Rongo R, Alstergren P, Ammendola L, Bucci R, Alessio M, D'Antò V, et al. Temporomandibular joint damage in juvenile idiopathic arthritis: Diagnostic validity of diagnostic criteria for temporomandibular disorders. *J Oral Rehabil* 2019.

12. Bernini JM, Kellenberger CJ, Eichenberger M, Eliades T, Papageorgiou SN, Patcas R. Quantitative analysis of facial asymmetry based on three-dimensional photography: a valuable indicator for asymmetrical temporomandibular joint affection in juvenile idiopathic arthritis

patients? *Pediatr Rheumatol Online J* 2020;18. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6995089/. Accessed June 25, 2020.

13. Zwir LF, Terreri MT, Amaral e Castro A do, Rodrigues WDR, Fernandes ARC. Is power Doppler ultrasound useful to evaluate temporomandibular joint inflammatory activity in juvenile idiopathic arthritis? *Clin Rheumatol* 2020;39:1237–1240.

14. Tolend MA, Twilt M, Cron RQ, Tzaribachev N, Guleria S, Kalle T von, et al. Toward Establishing a Standardized Magnetic Resonance Imaging Scoring System for Temporomandibular Joints in Juvenile Idiopathic Arthritis. *Arthritis Care Res* 2018;70:758–767.

15. Stoll ML, Guleria S, Mannion ML, Young DW, Royal SA, Cron RQ, et al. Defining the normal appearance of the temporomandibular joints by magnetic resonance imaging with contrast: a comparative study of children with and without juvenile idiopathic arthritis. *Pediatr Rheumatol Online J* 2018;16. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5784616/. Accessed April 27, 2018.

16. Ma GMY, Calabrese CE, Donohue T, Peacock ZS, Caruso P, Kaban LB, et al. Imaging of the Temporomandibular Joint in Juvenile Idiopathic Arthritis: How Does Quantitative Compare to Semiquantitative MRI Scoring?. *J Oral Maxillofac Surg* 2019;77:951–958.

17. Tzaribachev N, Fritz J, Horger M. Spectrum of magnetic resonance imaging appearances of juvenile temporomandibular joints (TMJ) in non-rheumatic children. *Acta Radiol Stockh Swed 1987* 2009;50:1182–6.

18. Ryan M, Farrar S. Using conjoint analysis to elicit preferences for health care. *BMJ* 2000;320:1530–1533.

19. Burnett HF, Regier DA, Feldman BM, Miller FA, Ungar WJ. Parents' preferences for drug treatments in juvenile idiopathic arthritis: a discrete choice experiment. *Arthritis Care Res* 2012;64:1382–1391.

20. Neogi T, Aletaha D, Silman AJ, Naden RL, Felson DT, Aggarwal R, et al. The 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for rheumatoid arthritis: Phase 2 methodological report. *Arthritis Rheum* 2010;62:2582–2591.

21. Krantz DH. Measurement Structures and Psychological Laws. *Science* 1972;175:1427–1435.

22. Hansen P, Ombler F. A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. *J Multi-Criteria Decis Anal* 2008;15:87–107.

23. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–745.

24. Kuno H, Jara H, Buch K, Qureshi MM, Chapman MN, Sakai O. Global and Regional Brain Assessment with Quantitative MR Imaging in Patients with Prior Exposure to Linear Gadolinium-based Contrast Agents. *Radiology* 2016;283:195–204.

25. Murata N, Murata K, Gonzalez-Cuyar LF, Maravilla KR. Gadolinium tissue deposition in brain and bone. *Magn Reson Imaging* 2016;34:1359–1365.

26. Vaid YN, Dunnavant FD, Royal SA, Beukelman T, Stoll ML, Cron RQ. Imaging of the temporomandibular joint in juvenile idiopathic arthritis. *Arthritis Care Res* 2014;66:47–54.

27. Koos B, Tzaribachev N, Bott S, Ciesielski R, Godt A. Classification of temporomandibular joint erosion, arthritis, and inflammation in patients with juvenile idiopathic arthritis. *J Orofac Orthop* 2013;74:506–19.

28. Kellenberger CJ, Arvidsson LZ, Larheim TA. Magnetic resonance imaging of temporomandibular joints in juvenile idiopathic arthritis. *Semin Orthod* 2015;21:111–120.

29. Kellenberger CJ, Junhasavasdikul T, Tolend M, Doria AS. Temporomandibular joint atlas for detection and grading of juvenile idiopathic arthritis involvement by magnetic resonance imaging. *Pediatr Radiol* 2018;48:411–426.

Table 1: Discrete choice experiment-derived relative importance weights for the items and levels of the JAMRIS-TMJ.

A) Inflammatory Domain

		Bone Marrow Edema	%	Bone Marrow Enhancement	%	Effusion	%	Joint Enhancement	%	Synovial Thickening	%
Grading Level	0	Absent	0	Absent	0	Normal: ≤1mm in the largest joint recess	0	Normal: No exceeding joint enhancement	0	Normal: no synovium visible	0
	1	Present	9	Present	10	Mild: >1 and ≤2mm in the largest joint recess	8	Mild: localized exceeding joint enhancement	17	Mild: ≤2mm thickness at the point of maximum synovial thickening	15
	2					Moderate/Severe: >2mm focally and/or extension to entire joint	16	Moderate/Severe: exceeding joint enhancement diffusely involving the joint	34	Moderate/Severe: >2mm	31

B) Damage Domain

		Condylar Flattening	%	Erosions	%	Disk Abnormalities	%
Grading Level	0	Normal round/ovoid shape	0	No irregularities or deep breaks	0	Absent	0
	1	Mild: Extent of flattening involves part of the surface of the condyle	17	Mild: Presence of irregularities involving only part of the articular surface of the condyle	21	Present	13
	2	Moderate/Severe: Extent of flattening involves the entire surface of the condyle, or loss of height in the condyle	38	Moderate/Severe: Presence of deep breaks in the subchondral bone seen in two planes, or irregularities involving the entire articular surface of the condyle	49		

Legend: After an image has been graded, the total score for each domain is calculated by adding the percentage weight of each given grade for all items to yield a scaled percentage disease severity score ranging from 0-100% for each domain separately. Weights presented in this table are the group-averaged weights from Figures 2 and 3.

Figures and Legends:

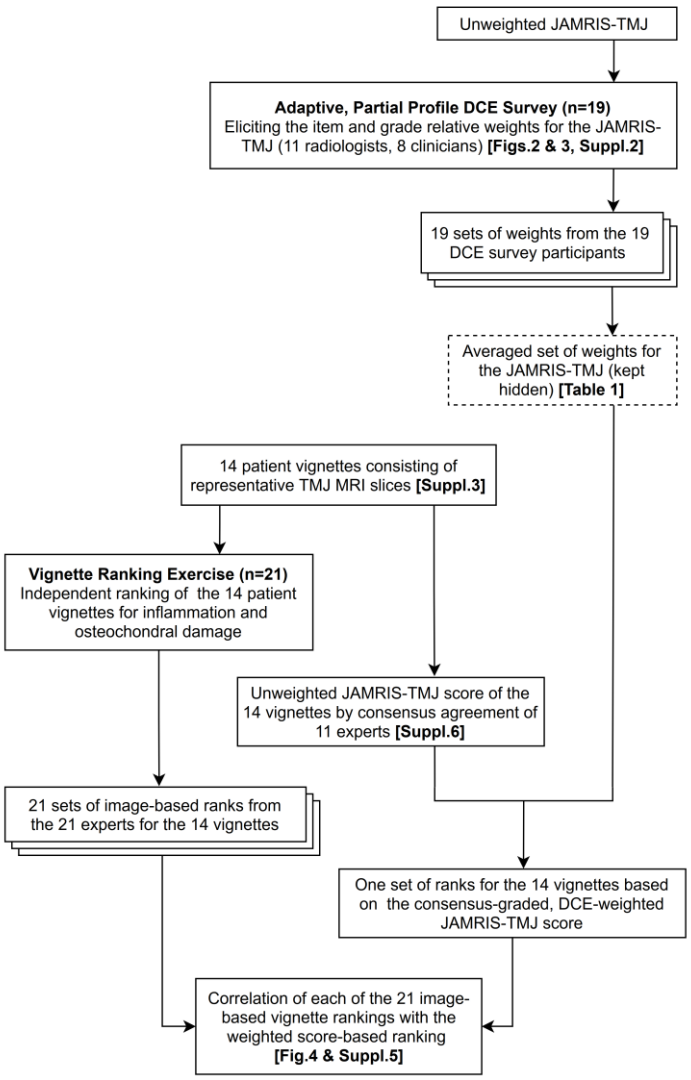


Figure 1: Flowchart summary outlining the progression of the study tasks in chronological order from top to bottom. First, an adaptive partial profile DCE survey was completed individually by a group of experts (n=19) to determine the importance weights of the items and grades of JAMRIS-TMJ. Second, blinded to the DCE-derived weights, an image-based vignette ranking exercise was completed individually by experts (n=21), producing 21 sets of both the inflammatory disease and osteochondral damage severity rankings for a set of 14 patient vignettes, based on full-profile, scoring system-independent method of comparison. Then, the item-wise JAMRIS-TMJ grades for the vignettes were agreed upon by consensus of experts (n=11), and the DCE-derived weights were applied to obtain the consensus weighted score for the vignettes for the two domains. Finally, the resulting vignette rankings from the two methods were correlated to test for convergent validity of the weighted JAMRIS-TMJ. Abbreviations: DCE, discrete choice experiment; JAMRIS-TMJ, Juvenile Idiopathic Arthritis Magnetic Resonance Imaging Scoring System for temporomandibular joints; Fig., figure in manuscript; Suppl., supplementary file available online.

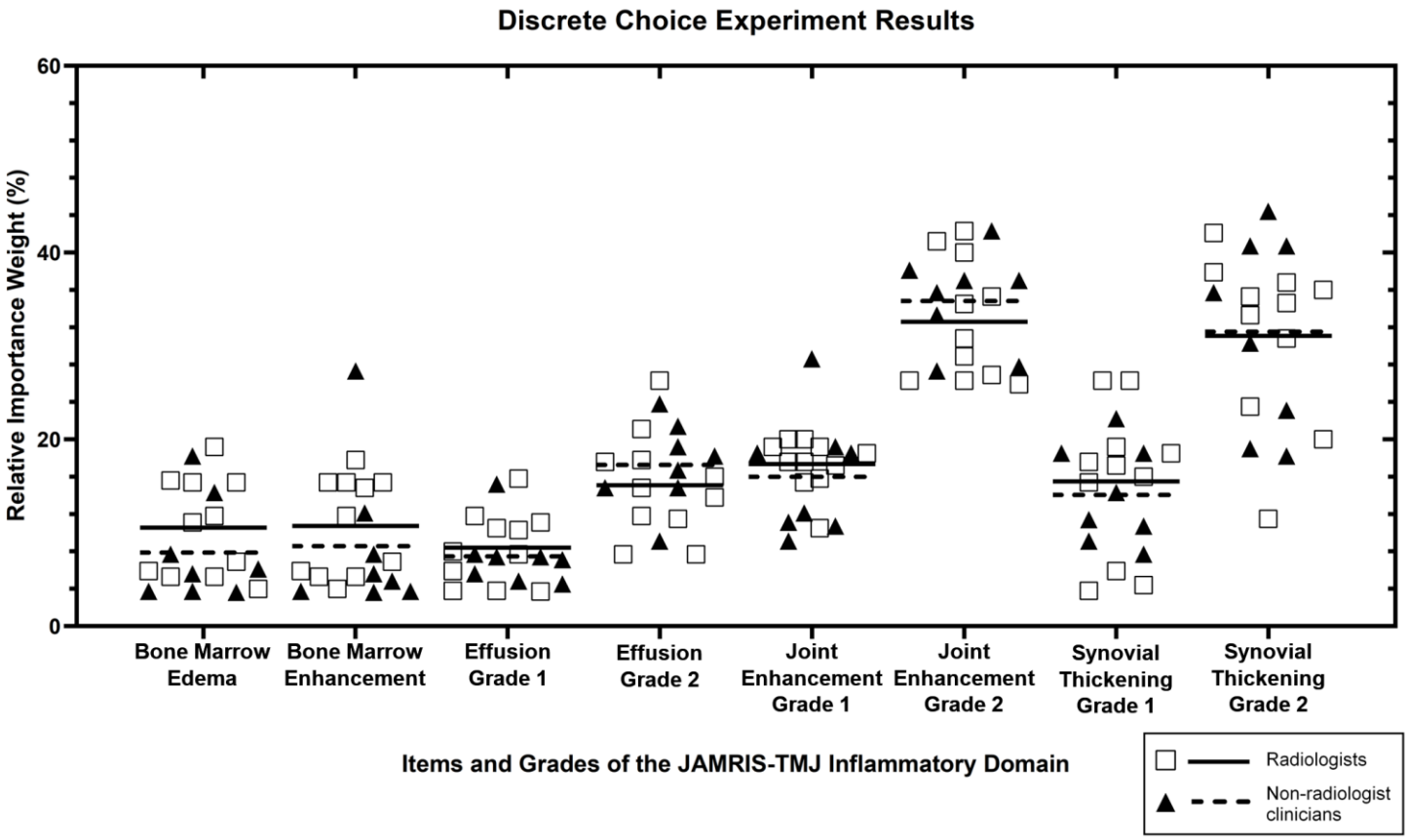


Figure 2: Scatterplot of the item and grade relative weights obtained from the discrete choice experiment survey for the JAMRIS-TMJ inflammatory domain. Relative importance weights from each of the participants are plotted, with lines indicating the average weight for radiologists (square marker and solid line, n=11) and non-radiologist clinicians (triangle marker and broken line, n=8) for each of the item-grades.

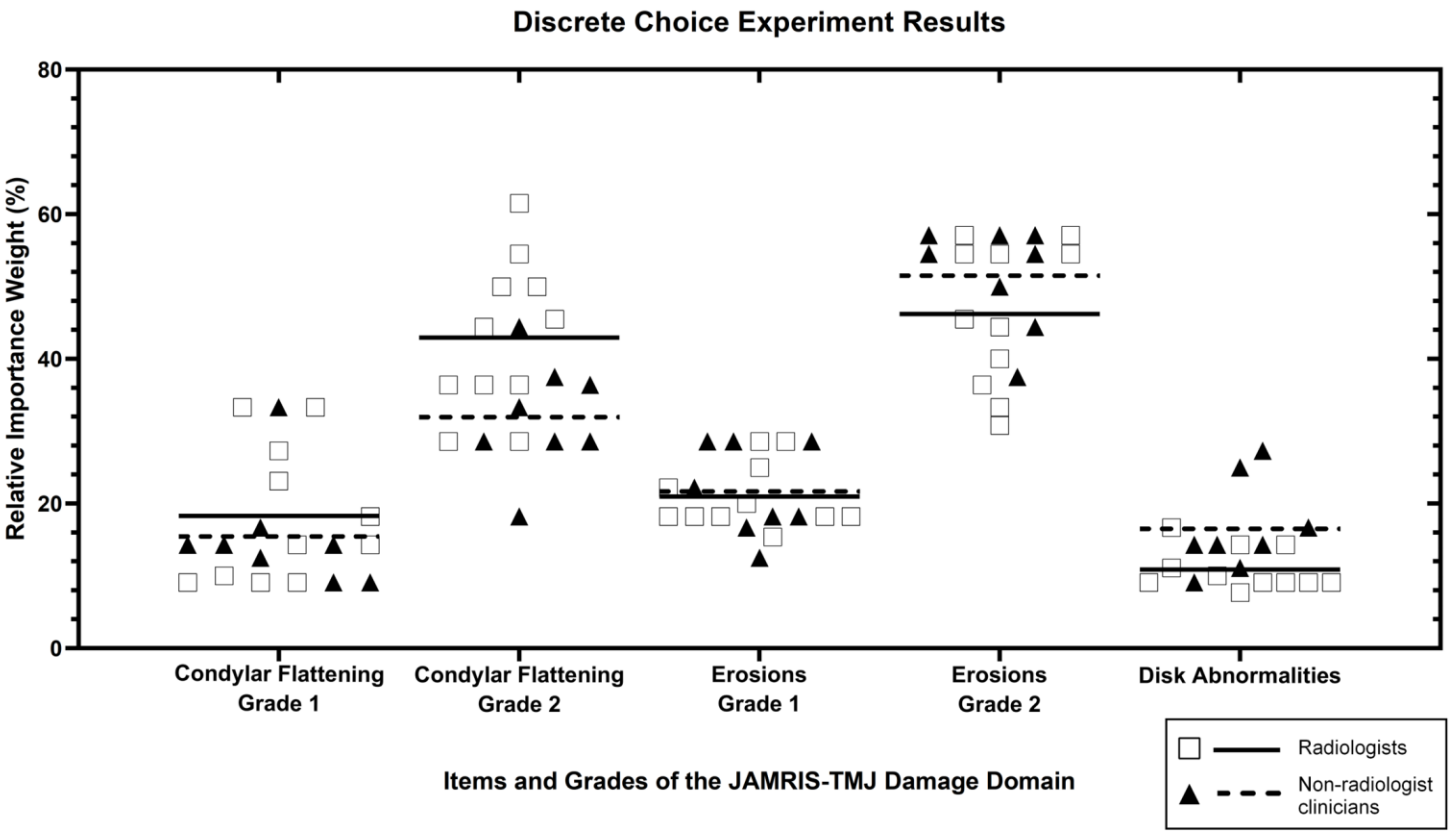


Figure 3: Scatterplot of the item and grade relative weights obtained from the discrete choice experiment survey for the JAMRIS-TMJ damage domains. Relative importance weights from each of the participants are plotted, with lines indicating the average weight for radiologists (square marker and solid line, n=11) and non-radiologist clinicians (triangle marker and broken line, n=8) for each of the item-grades.

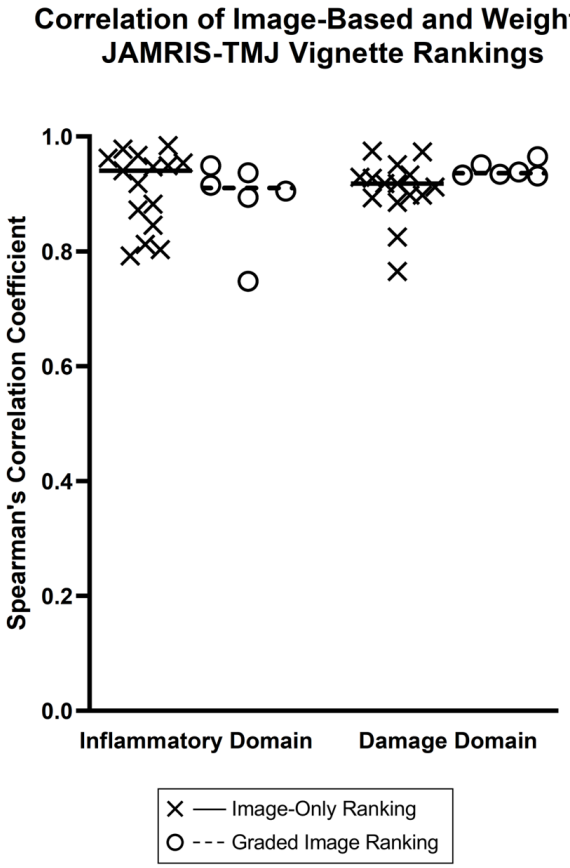


Figure 4: Correlation of the vignette ranks produced by the unrestricted, image-based ranking method and the weighted JAMRIS-TMJ score. Separate Spearman's rank correlation coefficients (rho) are plotted for each of the participants (n=21 total), comparing their image-based ranking of the 14 patient vignettes with one set of consensus-graded, DCE-weighted JAMRIS-TMJ score ranks for the vignettes. Horizontal lines represent the median Spearman's rho for each subgroup of participants: one group ranked the vignettes by the images only (cross marker and solid line, n=15), whereas the other group consisting of pediatric rheumatologists who do not regularly interpret and grade TMJ MRI exams themselves ranked the vignettes by the unweighted, pre-graded images (circle marker and broken line, n=6). The data which these coefficients derive from are visualized as scatterplots in Supplementary file 5.

Supplementary Files:

Supplementary file 1: Additional background information on discrete choice experiments as related to this study.

A discrete choice experiment (DCE) is characterized by questions that ask the participants to compare multiple attributes of a construct conjointly, applying trade-offs between these features according to the participant's opinion on their relative importance weights. DCE techniques are frequently used in choice modeling scenarios in economics and marketing, including the quantification of the relative importance of attributes in a multi-attribute entity (product, job offer, policy, etc.), determining the thresholds for trade-offs between the different attributes, and estimating the overall utility or probability of take-up of a multi-attribute entity. The technique is also becoming increasingly common in healthcare research where it is used to elicit and quantify the opinions of patients and medical experts on prioritizing treatment delivery, determining stakeholder preferences (1), and developing disease classification criteria and outcome measures (2).

The choice tasks in DCE may represent each alternative using part or all of its attributes, corresponding to partial-profile or full-profile designs, respectively. When the number of attributes is high, full-profile comparisons are cognitively more difficult to perform, which may lead the participants to give inconsistent responses, or cause them to use or develop simplifying heuristics by rating only based on a few important attributes while disregarding the rest. On the other hand, partial-profile comparisons rely on the assumption of single- and joint-factor independence, which state that the relative ordering of the levels within and between attributes, respectively, do not vary depending on the level of an external attribute (3). This assumption

may not always hold in the assessment of TMJ MRI. For example, synovial thickening may be considered more important than bone marrow changes and joint effusion if the thickened portion of synovium also enhances post-contrast, compared to when it does not. Joint effusion and synovial enhancement may be considered less important than other items when the condyle is severely damaged, as these features can be confounded by the mechanical irritation. Therefore, although it is more practical to conduct, the weights derived from a partial-profile DCE method should be validated, at least at face value, by comparing with another approach that utilizes full-profile comparisons.

A popular option for partial-profile DCE in rheumatology literature is the software developed by 1000Minds (Wellington, New Zealand) (4). An advantage of the 1000Minds DCE method compared to other partial-profile DCE designs is that as the respondent answers each question, the software adaptively reduces a large proportion of the potential comparison questions that are necessary to determine the full set of relative weights for the scoring system. While non-adaptive DCE designs will typically require a large sample of respondents to divide the necessary comparison questions, the adaptive 1000Minds method can derive the full set of weights from just one respondent, while still keeping the choices easy to compare and number of questions manageable. To achieve this efficiency, in addition to the factor independence assumption described above, this program also utilizes the transitive conservation of item-grade relationships (i.e., if A>B and B>C, then A>C) to implicitly solve a large proportion of the necessary discrete choice comparisons based on the participant's prior response. However, the transitivity assumption can further limit the capturing of more nuanced synergistic item relationships in the DCE results, depending on whether such a comparison is answered explicitly by the participants, or solved implicitly by the software. Furthermore, the adaptive nature of the 1000Minds DCE

can inflate the impact of respondent error, such as from misunderstanding the terms, cognitively difficult comparisons, respondent fatigue, or simplifying heuristics, since decision on every comparison question is formulaically applied to solve and reduce a large number of potentially necessary comparisons. Therefore, in this study, we tested the relative weights produced from the adaptive partial-profile DCE against a more holistic approach utilizing full-profile vignette-to-vignette comparison that avoids these reductionistic assumptions.

References

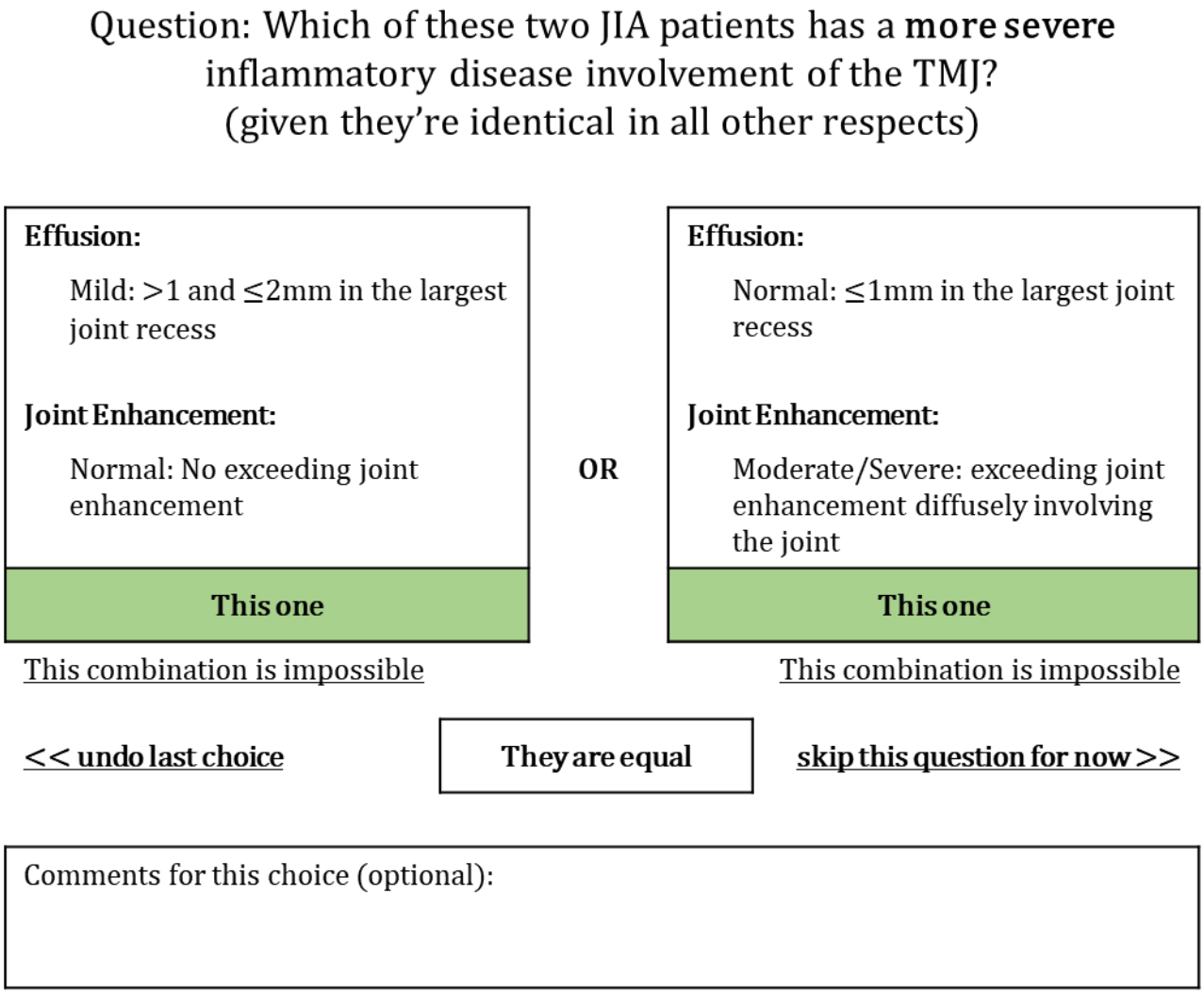
1. Burnett HF, Regier DA, Feldman BM, Miller FA, Ungar WJ. Parents' preferences for drug treatments in juvenile idiopathic arthritis: a discrete choice experiment. *Arthritis Care Res* 2012;64:1382–1391.

2. Neogi T, Aletaha D, Silman AJ, Naden RL, Felson DT, Aggarwal R, et al. The 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for rheumatoid arthritis: Phase 2 methodological report. *Arthritis Rheum* 2010;62:2582–2591.

3. Krantz DH. Measurement Structures and Psychological Laws. *Science* 1972;175:1427–1435.

4. Hansen P, Ombler F. A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. *J Multi-Criteria Decis Anal* 2008;15:87–107.

Supplementary file 2: Example of a discrete choice comparison question that the experts completed through the 1000Minds survey software. Separate DCE surveys were completed for the inflammation and osteochondral damage severities to weigh the JAMRIS-TMJ inflammatory and damage domains, respectively. Depending on the randomized question order and the participant response, approximately 20-25 questions in total were required from each participant to yield a complete set of relative weights for the two domains of JAMRIS-TMJ.



Supplementary file 3: Descriptions and rationales for the statistical tests used in this study.

Adaptive DCE survey to determine the importance weights: The 1000Minds DCE platform which we used for eliciting and quantifying the opinions of the experts on importance weights is based on a technique called Potentially All Pairwise Rankings of all possible Alternatives (PAPRIKA). The method has been published and validated (1), and the 1000Minds software has been used in various commercial and scientific research settings, including recent studies in rheumatology on determining patient and expert preferences (2,3). We chose this method over other more traditional DCE designs for its adaptive nature that allows determining a complete set of relative weights for every respondent while keeping the discrete choice comparison questions relatively few and easy. Although the obtained weight estimates may not be as robust and generalizable as compared to other non-adaptive DCE designs with pre-determined set of questions administered to large number of experts, the 1000Minds DCE methods enabled us to achieve representative quantitative data despite having a limited sample of experts with domain expertise on this specialized outcome measurement topic.

Assessment of the representativeness of the average weight (Spearman's rho): The DCE survey provided a complete set of ratio-level item- and grade-wise importance weight data for each of the 19 participants in the DCE exercise. Thus, there were 19 numbers for each grade of every item in the two scoring system domains (8 item-grades for the inflammatory, and 5 item-grades for the damage domain). The average of the 19 weights yields a 20^{th} set of group-averaged weights per each item-grade. How well this averaged set of weights represents the group homogeneity is assessed by correlating the ranking of all potential item combinations produced by every respondent's unique set of item-grade weights to the rank produced by the group-

averaged set of weights. For example, in the inflammatory domain, there are five items with 2 or 3 levels each, resulting in 108 potential item combinations, or profiles (2x2x3x3x3). These 108 profiles are ranked from least to greatest score in 20 different orders: 19 orders using each of the 19 respondents' personal weights, and a 20th order representing the group-averaged weight. Spearman's rank correlation coefficient (rho) was used to correlate each of the 19 respondents' rank order to the group-averaged order for each of these 108 item combinations for the inflammatory domain. Since there are 19-individual vs 1-averaged comparisons done pairwise, there are 19 rho coefficients, presented as median, IQR and min/max. Although this method seems indirect and somewhat complicated, it is the default metric calculated by the 1000Minds platform for assessing homogeneity since it is consistent with how the PAPRIKA method derives the weights: the DCE survey adaptively asks or solves all the necessary questions until the respondents' ranking for all 108 profiles can be determined either explicitly or implicitly. For further details, please refer to the primary reference (1).

Assessment of homogeneity of elicited weights (ICC): The agreement of the 19 sets of importance weights from the 19 experts across the 8 inflammatory domain item-grades was calculated by the intraclass correlation coefficient (ICC), using the single-measure, two-way random, absolute agreement definition [ICC(2,1)]. The ICC was used because it is a familiar coefficient of agreement for ratio-level data. The ICC(2,1) definition was used because it treats the participant identity as random effect, making the result more generalizable to other clinician experts. The rho coefficient described above and this ICC are both used to assess opinion homogeneity. The rho is used for pairwise correlation of average vs individual preference, yielding 19 coefficients per domain, whereas the ICC calculates the 19-participant agreement directly, yielding one ICC per domain. We believe both are helpful since the former is directly

assess the appropriateness of the average weight (i.e., how far each participant deviates from the average), while the latter is a direct quantification of the strength of group-level clustering of the importance weights.

Vignette rank clustering among experts (ICC): The clustering of the ranks given to the 14 vignettes by the 21 participants in the vignette ranking exercise was tested by the ICC(2,1) definition, as with the assessment of the clustering of DCE-derived weights. Dataset was organized for calculating 21-participant agreement across the 14 vignette ranks for the inflammatory and damage domain separately.

Vignette ranking correlation (Spearman's rho): The 14 vignettes can be ordered in two ways. First is by simply comparing the images as a whole between vignettes, without restricting the interpretation to the grade intervals and item definitions of any specific scoring system. The 21 participants in the vignette ranking exercises provided 21 sets of these rankings for both the inflammation and osteochondral damage assessment (yielding 21 ranks for each of 14 vignettes in 2 domains). Second method is by using the weighted JAMRIS-TMJ scoring system, with the group-averaged weights derived from the previous DCE. Raw JAMRIS-TMJ score was determined for the 14 vignettes by the consensus of 11 radiologists, which was multiplied with the weights to produce 0-100% inflammation and damage domain scores for the 14 vignettes (one comparator rank for each of 14 vignettes in 2 domains). Spearman's rank correlation coefficient (rho) was used to correlate the rankings from the two sources, i.e., participant 1's imaged based ranks of the 14 vignettes vs the consensus weighted scores (rho_1), participant 2's image-based vignette ranks to the same consensus weighted score (rho_2), participant 3's vs consensus weighted score (rho_3)… etc. for all 21 participants. The spread in the resulting 21 rho

coefficients are displayed in Figure 4. The raw sets of data that yield these 21 coefficients are visualized in Supplementary File 5 as scatterplots.

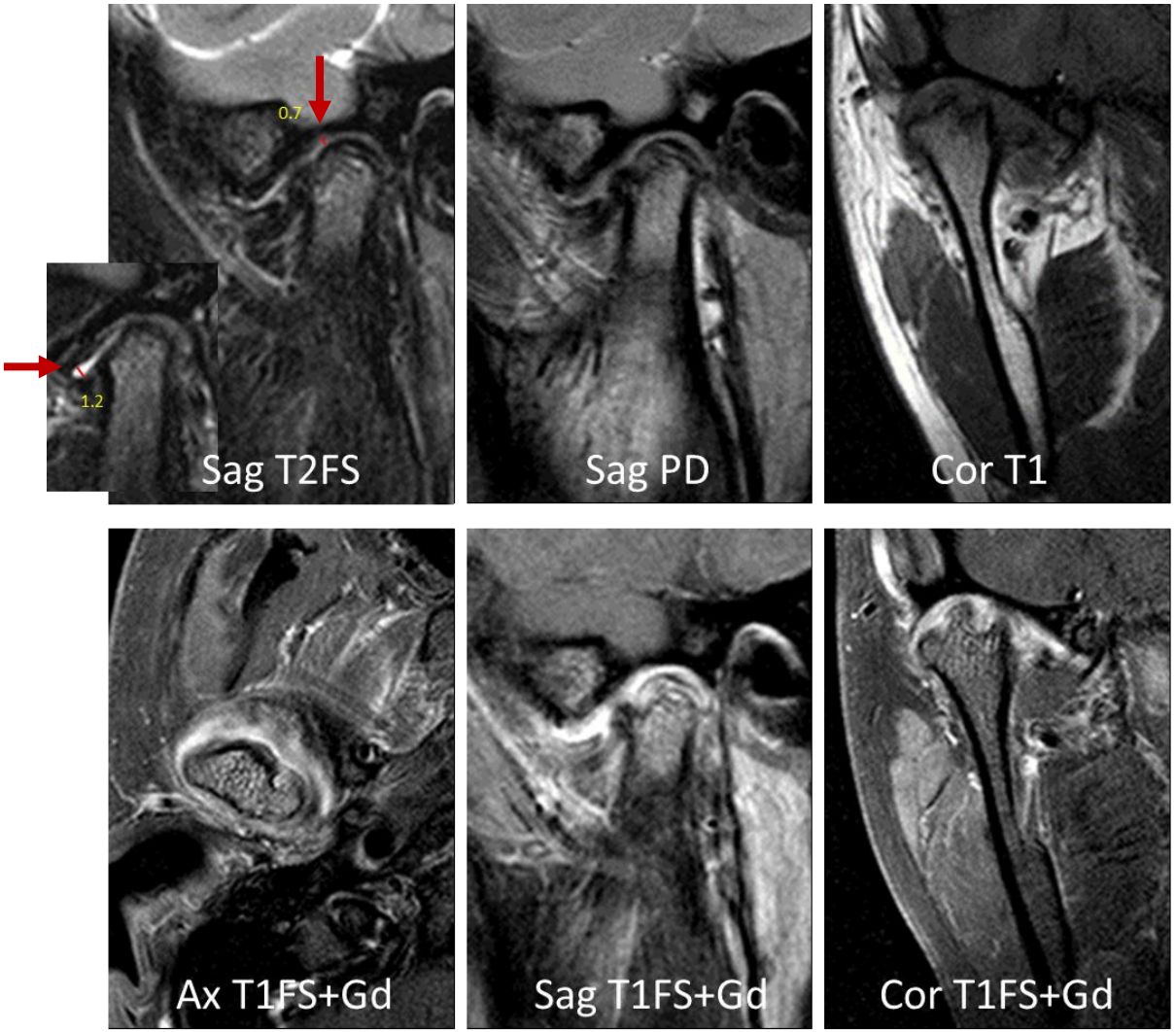
References

1. Hansen P, Ombler F. A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. *J Multi-Criteria Decis Anal* 2008;15:87–107.

2. Neogi T, Aletaha D, Silman AJ, Naden RL, Felson DT, Aggarwal R, et al. The 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for rheumatoid arthritis: Phase 2 methodological report. *Arthritis Rheum* 2010;62:2582–2591.

3. Burnett HF, Regier DA, Feldman BM, Miller FA, Ungar WJ. Parents' preferences for drug treatments in juvenile idiopathic arthritis: a discrete choice experiment. *Arthritis Care Res* 2012;64:1382–1391.

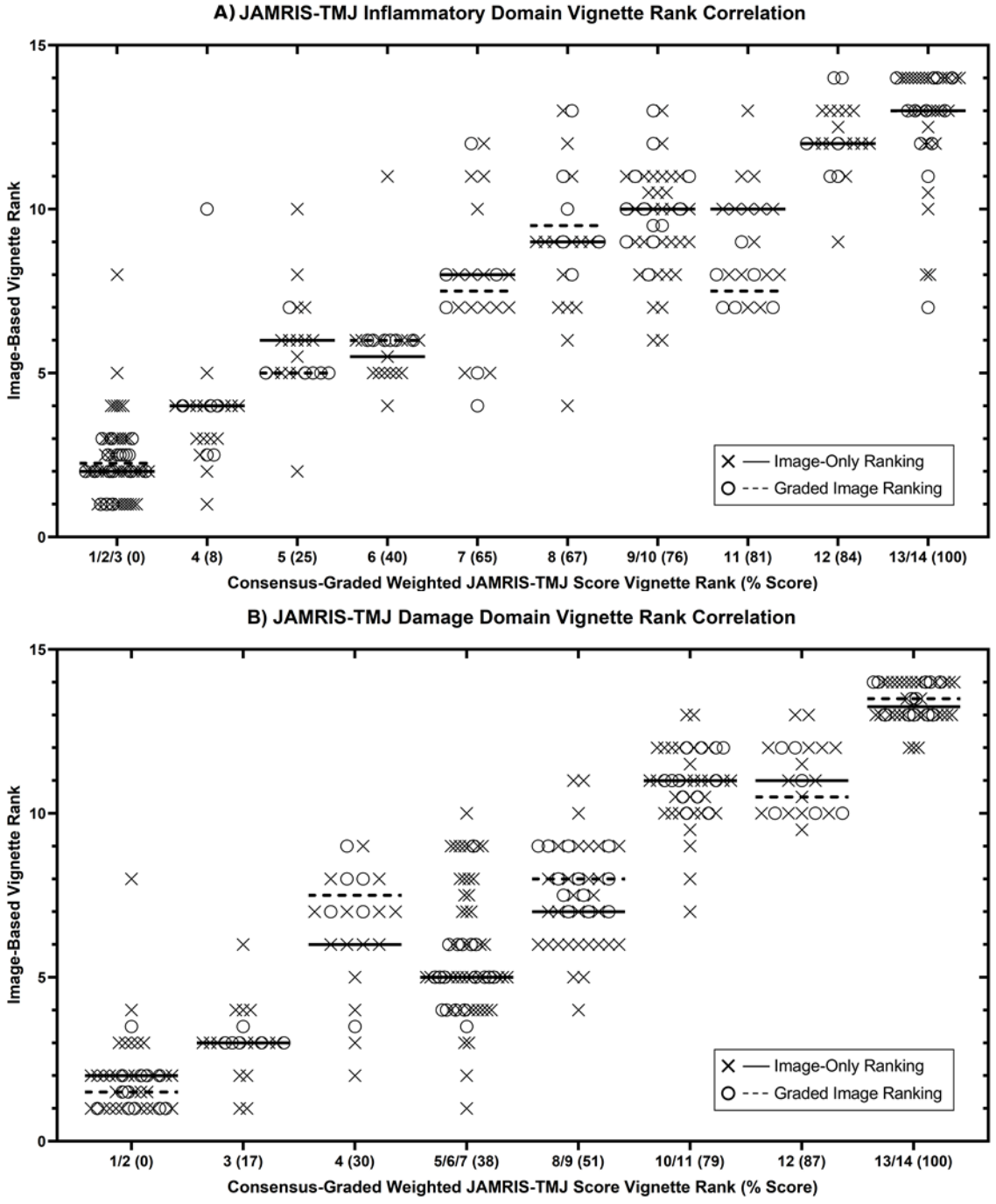
Supplementary file 4: Example of a temporomandibular joint MRI vignette ranked by the participants of the vignette ranking exercise. Image-based ranking of the vignettes was considered an unrestricted, scoring system-independent method for assessment of construct validity of the score scaling produced by the DCE survey. Participants in the graded vignette ranking subgroup referred to the "Radiologist's Assessment" grades, which are based on the JAMRIS-TMJ definitions. Abbreviations: Ax, axial; Cor, coronal; FS, fat suppression sequence; Gd, gadolinium; PD, proton density weighted sequence; Sag, sagittal. Red lines and arrows: measurements in mm of the anterior and mid synovium.



Case L

Radiologist's Assessment	
Inflammatory Domain	
Bone Marrow Edema (0-1):	1
Bone Marrow Enhancement (0-1):	1
Effusion (0-2):	1
Joint Enhancement (0-2):	2
Synovial Thickening (0-2):	1
Damage Domain	
Condylar flattening (0-2):	1
Erosions (0-2):	2
Disk Abnormalities (0-1):	1

Supplementary file 5: Scatterplots of the vignette ranks produced by the holistic, image-based method of ranking, ordered by the consensus-graded, weighted JAMRIS-TMJ score of the vignettes for the **A)** inflammatory and **B)** damage domains. Fifteen experts participated in the image-only ranking (cross marker), and 6 in the graded image ranking (circle marker). Each of the 14 vignettes has one x-axis value corresponding to its consensus weighted score rank, and 21 y-axis values for the image-based ranks given individually by the experts. Vignettes which received the same weighted score are plotted together as a group (/). Horizontal lines represent the median weighted score rank given to each of the image-based vignette ranks.



Supplementary file 6: Tables illustrating the item-wise breakdown of consensus grades for the 14 vignettes based on the JAMRIS-TMJ definitions (1). The 14 columns correspond to the 14 vignettes, ordered by their resulting weighted score-based ranks. The group-averaged weights from the DCE are applied to produce the weighted domain scores. Column headers for the ranks are combined for vignettes which received the same weighted score.

		Rank of Vignette Based on Weighted Score													
		2			4	5	6	7	8	9.5		11	12	13.5	
Inflammatory Domain Item	Bone Marrow Edema	0	0	0	0	0	0	0	1	1	1	0	1	1	1
	Bone Marrow Enhancement	0	0	0	0	0	0	0	1	1	1	0	1	1	1
	Effusion	0	0	0	1	1	1	0	2	1	1	2	2	2	2
	Joint Enhancement	0	0	0	0	1	1	2	1	2	2	2	2	2	2
	Synovial Thickening	0	0	0	0	0	1	2	1	1	1	2	1	2	2
	Raw Score	0	0	0	1	2	3	4	6	6	6	6	7	8	8
	Weighted Score (%)	0	0	0	8	25	40	65	67	76	76	81	84	100	100

		Rank of Vignette Based on Weighted Score													
		1.5		3	4	6			8.5		10.5		12	13.5	
Damage Domain Item	Condylar Flattening	0	0	1	1	1	1	1	1	1	1	1	2	2	2
	Erosions	0	0	0	0	1	1	1	1	1	2	2	2	2	2
	Disk Abnormalities	0	0	0	1	0	0	0	1	1	1	1	0	1	1
	Raw Score	0	0	1	2	2	2	2	3	3	4	4	4	5	5
	Weighted Score (%)	0	0	17	30	38	38	38	51	51	79	79	87	100	100

Reference:

1. Tolend MA, Twilt M, Cron RQ, Tzaribachev N, Guleria S, Kalle T von, et al. Toward Establishing a Standardized Magnetic Resonance Imaging Scoring System for Temporomandibular Joints in Juvenile Idiopathic Arthritis. *Arthritis Care Res* 2018;70:758–767.