Follow us @BoneJointOpen

Freely available online OPEN ACCESS

BJO

GENERAL ORTHOPAEDICS

Computerized adaptive testing for the Oxford Hip, Knee, Shoulder, and Elbow scores

ACCURATE MEASUREMENT FROM FEWER, AND MORE PATIENT-FOCUSED, QUESTIONS

C. J. Harrison,
O. R. Plummer,
J. Dawson,
C. Jenkinson,
A. Hunt,
J. N. Rodrigues

From Universal Research Solutions, Columbia, Missouri, USA

Aims

The aim of this study was to develop and evaluate machine-learning-based computerized adaptive tests (CATs) for the Oxford Hip Score (OHS), Oxford Knee Score (OKS), Oxford Shoulder Score (OSS), and the Oxford Elbow Score (OES) and its subscales.

Methods

We developed CAT algorithms for the OHS, OKS, OSS, overall OES, and each of the OES subscales, using responses to the full-length questionnaires and a machine-learning technique called regression tree learning. The algorithms were evaluated through a series of simulation studies, in which they aimed to predict respondents' full-length questionnaire scores from only a selection of their item responses. In each case, the total number of items used by the CAT algorithm was recorded and CAT scores were compared to full-length questionnaire scores by mean, SD, score distribution plots, Pearson's correlation coefficient, intraclass correlation (ICC), and the Bland-Altman method. Differences between CAT scores and full-length questionnaire scores were contextualized through comparison to the instruments' minimal clinically important difference (MCID).

Results

The CAT algorithms accurately estimated 12-item questionnaire scores from between four and nine items. Scores followed a very similar distribution between CAT and full-length assessments, with the mean score difference ranging from 0.03 to 0.26 out of 48 points. Pearson's correlation coefficient and ICC were 0.98 for each 12-item scale and 0.95 or higher for the OES subscales. In over 95% of cases, a patient's CAT score was within five points of the full-length questionnaire score for each 12-item questionnaire.

Conclusion

Oxford Hip Score, Oxford Knee Score, Oxford Shoulder Score, and Oxford Elbow Score (including separate subscale scores) CATs all markedly reduce the burden of items to be completed without sacrificing score accuracy.

Cite this article: *Bone Jt Open* 2022;3-10:786–794.

Keywords: Patient-reported outcome measures, Computerized adaptive testing, Machine-learning, Oxford Hip Score, Oxford Knee Score, Oxford Shoulder Score

Introduction

Patient-reported outcome measures (PROMs) are questionnaires used to measure health constructs, like knee, hip, elbow, or shoulder pain and function, or quality of life, in clinical practice and research. The subjective element within these constructs cannot be quantified directly, hence the importance placed on PROMs. Recognition of this has led to PROMs being used in large

Correspondence should be sent to Conrad J Harrison; email: team@methodologyoxford.com

doi: 10.1302/2633-1462.310.BJO-2022-0073.R1

Bone Jt Open 2022;3-10:786–794.

initiatives such as the UK NHS PROMs Programme, which has captured data for hundreds of thousands of procedures undertaken since April 2009.[1] PROMs are widely used as outcome measures in high-quality randomized controlled trials, and there are specific guidelines for their use.[2] It is possible that PROM use may go beyond measurement, and actually improve clinical outcomes, by affecting communication and flagging problems.[3,4]

Well-designed PROMs are available for orthopaedic and musculoskeletal conditions. These include the Oxford Hip Score (OHS),[5,6] Oxford Knee Score (OKS),[7] Oxford Shoulder Score (OSS),[8] and Oxford Elbow Score (OES).[9] There are key strengths to this group of PROMs; in particular, patients were centrally involved in their development. All items were derived from patients' interview accounts and scrutinized by clinicians. Consequently, they comprise items that matter to patients as well as to clinicians. Their high quality has contributed to their being used widely in studies, and in initiatives like the NHS PROMs programme, as well as in other registry systems around the world.[10]

Despite the theoretical importance of PROMs, and the widespread availability of questionnaires for many clinical scenarios, PROMs have not fulfilled their potential in clinical practice.[11] Recent in-depth National Institute for Health Research (NIHR)-funded work has highlighted several strains that may contribute to this.[12] A key system strain is the trade-off between the high burden of completing full-length PROMs that are well-validated, versus using shorter measures that may be more practical in real-world use but risk sacrificing validity.[12] Full-length and fixed shortened PROMs may not always relate to individual patients' experiences.[12]

Computerized adaptive testing (CAT) can provide a solution. Rather than pose every item in a PROM, CATs use patterns in people's responses to select the most appropriate items to administer to a person, with not all items required to estimate a person's score. This is analogous to skipping questions that are not relevant when taking a history. For example, somebody who says they cannot jog does not need to be asked if they can sprint. CATs can shorten PROMs, making them easier to deploy in practice, while maintaining the originally validated items and better tailor them to the individual patient by selecting questions that will provide the most information. We developed machine-learning-based CATs for the OHS, OKS, OSS, and OES using the OBERD software system (Universal Research Solutions, USA), which collects outcome measure data in clinical practice.

Methods

We used a form of supervised machine-learning known as regression tree learning to develop CAT assessments for the Oxford Hip, Knee, Shoulder, and Elbow Scores.[13] Presented with a large dataset of respondents' scores to each item in a PROM, these algorithms aim to iteratively split respondents into groups with similar overall scores, based on their responses to individual items. The algorithms learn how to split the data in the most efficient way possible, i.e. which items to pick, and in which order, to create subgroups of respondents with scores that are as similar as possible, from as few item responses as possible. These data splits can then be applied to select items in a CAT assessment. For example, the OHS algorithm might start by asking "How would you describe the pain you usually have in your hip?" and a respondent might select "None". The algorithm may demonstrate that on average, the group of people who select that response go on to score a total of 45 on the OHS. At that point the algorithm will impute a respondent score of 45, or pose another item to split that group into further subgroups with narrower score distributions (for a more accurate estimation).

For each PROM, we developed CAT algorithms from retrospective datasets that included deidentified patient responses to the respective full-length questionnaire. These were collected during routine clinical practice from multiple clinical practices, using tablets or laptops which accessed the internet-based OBERD system (USA), and included both preoperative and postoperative responses from patients undergoing clinical evaluation for any condition on an inpatient or outpatient basis.

We randomly split each dataset into training and test sets without data leakage, and used the training sets to develop the CAT algorithms. This aimed to ensure that preoperative and postoperative forms were included in each. The approach was to assign about 80% for training and 20% for testing. However, after the models were finalized, additional data became available for OKS and OHS and these were added to the test set in order to obtain the most comprehensive possible evaluation. The exact counts are shown in Table I. The percentages are not exact because a few forms had to be dropped after the randomization, due to missing responses. The specific choice of the first item posed by our CATs was derived from the training set by the learning algorithms as the single most explanatory question.

We then used the test sets (Table I) to evaluate the performance of each algorithm in terms of its accuracy (correlation and concordance of CAT score and full-length questionnaire score) and its ability to reduce the number of items administered. To do this, we programmed CAT simulation experiments in which CAT algorithms aimed to reproduce a respondent's total scale score using only a selection of individual item responses. The CAT algorithms were free to pick which items to use, and in which order (as they would do in a real-life setting). In each instance, we recorded the total number of items used by the CAT algorithm and the overall score estimate.

Table I. Demographic data for each of the test sets, as well as the number of physicians who contributed patient responses to these datasets.

Variable	OKS	OHS	OSS	OES
Sample size	5,622	3,471	561	2,084
Physicians contributing data, n	7	6	18	31
Median age, yrs (IQR)	68 (62 to 74)	67 (60 to 73)	66 (54 to 74)	56 (46 to 65)
Age group, n (%)				
≤ 30 yrs	5 (0.09)	10 (0.29)	43 (7.66)	195 (9.36)
31 to 45 yrs	72 (1.28)	101 (2.91)	40 (7.13)	311 (14.9)
46 to 60 yrs	1,059 (18.8)	851 (24.5)	127 (22.6)	776 (37.2)
61 to 75 yrs	3,358 (59.7)	1,918 (55.3)	241 (43)	690 (33.1)
> 75 yrs	1,128 (20.1)	591 (17)	108 (19.2)	108 (5.18)
Unknown	0	0	2 (0.36)	4 (0.19)
Sex, n (%)				
Female	3,778 (67.2)	1,984 (57.2)	275 (49.02)	1,015 (48.7)
Male	1,843 (32.8)	1,486 (42.8)	245 (43.67)	1,064 (51.06)
Unknown	1 (0.02)	1 (0.03)	41 (7.31)	5 (0.24)

IQR, interquartile range; OES, Oxford Elbow Score; OHS, Oxford Hip Score; OKS, Oxford Knee Score; OSS, Oxford Shoulder Score.

To evaluate the suitability of CAT for population-level assessments, we compared CAT and full-length questionnaire scores with the following methods: means and standard deviation (SD), Pearson's correlation coefficient, intraclass correlation (ICC), and frequency distribution plots. To evaluate the suitability of CAT for supporting individual-level assessment, we compared CAT and full-length questionnaire with the Bland-Altman method.[14] This combination of analyses has been used previously, and considers the relationship between the CAT score and full length score, the contribution of inherent variability of the PROM itself to the apparent difference between CAT and full-length score, the relationship between the CAT and full-length score throughout the distribution of scores, and the pattern of differences between them.[15]

Differences between CAT and full-length scores were contextualized through comparison with the instruments' minimal clinically important difference (MCID, the smallest amount of change in a score that could be considered to have clinical importance).[16] This approach was chosen as MCID values are available for these PROMs. MCIDs are population-based, but as they are anchored to individual perception of meaningful difference, we considered them reasonable to use as a real-world interpretive guide only. Notably, the model development and accuracy assessment were performed separately from this. The MCID is context-specific, so we used MCID values from comparable clinical scenarios.

While the OHS, OKS, and OSS are scored and reported consistently within the literature, variation exists in the presentation of results from the OES. The OES was originally described as a multidimensional instrument comprising three unidimensional subscales (four items each) which measure elbow function, pain, and social-psychological impact.[9] The scales can be used individually for discrete measurement in each domain, or a total, multidimensional score can be obtained by combining the responses of all three subscales. Scores can be transformed into a 0 to 100 format, but are also often reported as raw sum scores, which range from 0 to 16 for each subscale or 0 to 48 for the overall score. We developed a single CAT for the OES to calculate scores in each of the three subscales and an overall score. For consistency with the other instruments, we report OES CAT results for the whole scale and subscales as raw sum scores.

Results

Oxford Hip Score. We included 8,471 OHS response sets, of which 5,000 were randomly allocated to the training dataset and 3,471 were allocated to the test dataset. To estimate full-length (12-item) OHS score, the CAT algorithm would have asked 13% of these respondents six questions, 76% of these respondents seven questions, and 11% of these respondents eight questions.

Full-length questionnaire scores and CAT scores followed a very similar distribution (Figure 1, Table II) with a difference in mean score of 0.22 out of 48 points. Pearson's correlation coefficient and ICC between CAT and full-length questionnaire scores were both 0.98 (Figure 2).

For 3,520 patients for whom both preoperative and postoperative scores were available, the mean difference in change score calculated from full-length OHS and CAT OHS was 0.35 points (SD 2.68), and the median difference in change score was 0.40 points (IQR -1.33 to 2.04).

Oxford Knee Score. We included 15,106 response sets to the 12-item OKS: 9,484 in the OKS training dataset and 5,622 in the corresponding test set. In 24% of cases the CAT algorithm used six items, in 58% of cases it used seven items, in 9% of cases it used eight items, and in 10% of cases it used nine items.

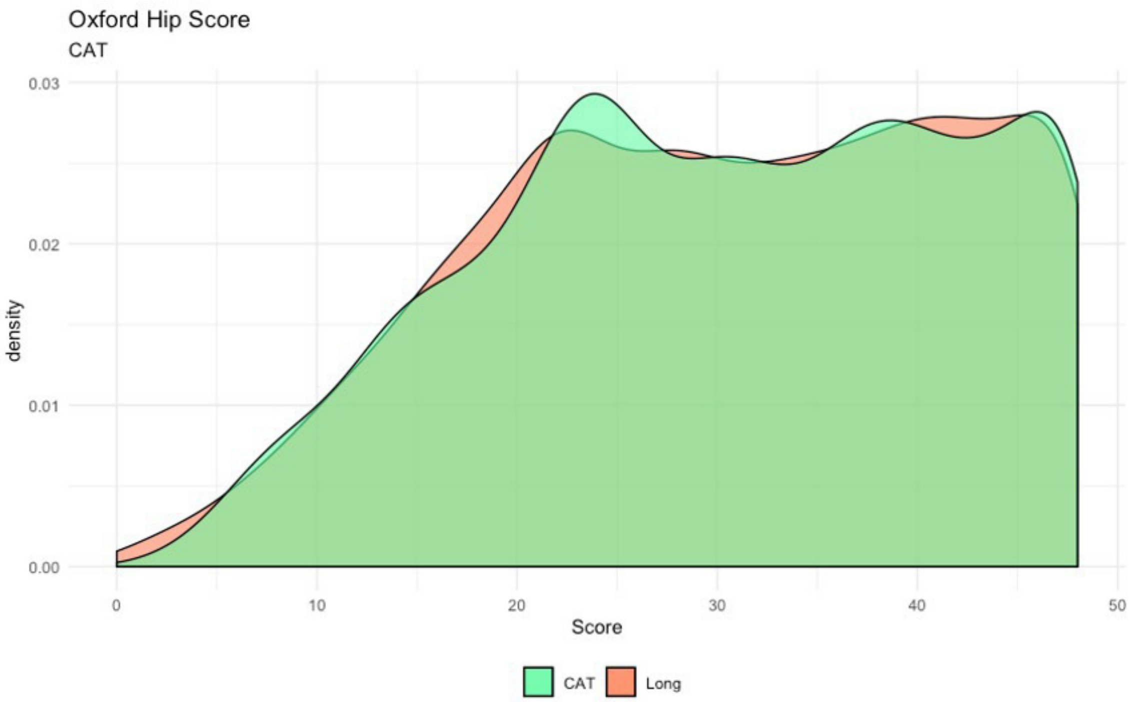

Fig. 1

Score distributions for the full-length Oxford Hip Score and computerized adaptive testing (CAT) version of the Oxford Hip Score.

Table II. Results from computerized adaptive testing simulation studies. Minimum and maximum differences in score are calculated as full-length score minus computerized adaptive testing score.

Measure	Training data sample size	Test data sample size	Mean full-length score (SD)*	Mean CAT score (SD)*	MCID*[17,18]	Minimum difference in score	Maximum difference in score	Pearson's correlation coefficient	ICC	Mean items administered by CAT algorithm, n
OHS (12 items total)	5,000	3,471	30.43 (11.70)	30.65 (11.59)	11	- 7.83	12.07	0.98	0.98	6.98
OKS (12 items total)	9,484	5,622	28.16 (10.67)	28.43 (10.61)	9	- 10.10	8.62	0.98	0.98	7.05
OSS (12 items total)	2,258	561	31.49 (11.23)	31.52 (11.33)	6	- 7.70	8.44	0.98	0.98	7.25
OES total (12 items total)	8,359	2,084	35.49 (11.85)	35.62 (11.83)	7.5	-10.17	12.00	0.98	0.98	4.87
$\text{OES}_{\text{Function}}$ (4 items total)	8,334	2,078	12.74 (3.60)	12.84 (3.50)	1.5	-7.16	4.36	0.95	0.95	
OES_{Pain} (4 items total)	8,344	2,078	11.36 (4.48)	11.33 (4.41)	3	-4.73	5.32	0.97	0.97	
$\text{OES}_{\text{Social}}$ (4 items total)	8,338	2,078	11.39 (4.72)	11.38 (4.63)	3	-5.56	5.57	0.98	0.97	

*Raw score.

CAT, computerized adaptive testing; ICC, intraclass correlation; MCID, minimal clinically important difference; OES, Oxford Elbow Score; OHS, Oxford Hip Score; OKS, Oxford Knee Score; OSS, Oxford Shoulder Score; SD, standard deviation.

The difference in mean score between the full-length assessment and the CAT version was 0.26 out of 48 points. Pearson's correlation coefficient and ICC were both 0.98. For context, this is better than the Pearson's correlation coefficient of test/retest completion of the OKS in the original development paper, which was 0.92.[5] For 95% of patients (n = 5,341), the difference between the CAT and full score was less than 4.5 points, which is smaller than the MCID (see Table II).

Both preoperative and postoperative scores were available for 4,453 patients. Between the full-length OKS and the CAT, the change scores differed by a mean of 0.34 points (SD 2.6) and a median of 0.36 points (IQR -1.37 to 2.05).

Oxford Shoulder Score. We included 2,819 response sets to the 12-item OSS: 2,258 in the training dataset and 561 in the test dataset. In 4% of cases, the CAT algorithm used six items, in 68% of cases the CAT algorithm used seven

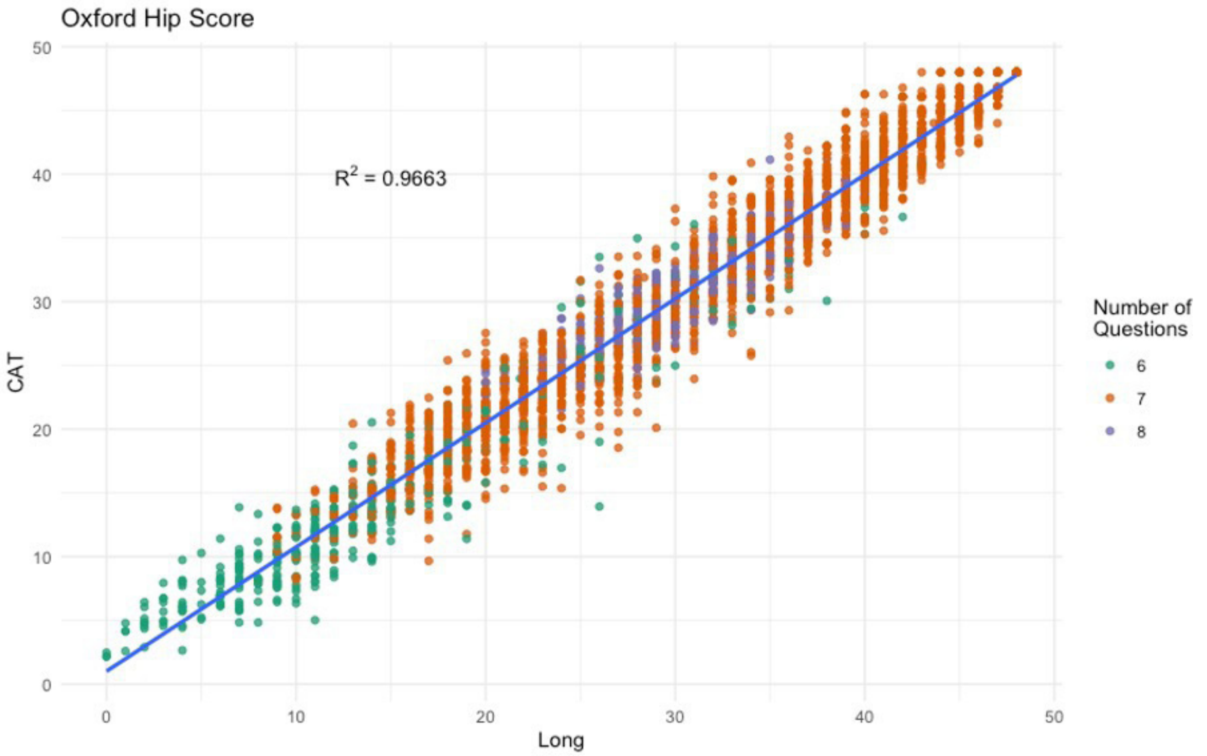

Fig. 2

Correlation of scores from the full-length (long) Oxford Hip Score and computerized adaptive testing (CAT) version of the Oxford Hip Score. Point colour indicates the number of questions used by the CAT algorithm in that instance.

items, and in 29% of cases the CAT algorithm used eight items.

The difference in mean score between the full-length OSS at the CAT version was 0.03 points out of 48 points. Pearson's correlation coefficient and ICC were 0.98.

In over 95% of cases (n = 533), CAT OSS scores and full-length OSS scores differed by under five points (Figure 3 and Figure 4), which is less than the six-point mean reported MCID for the instrument.[19]

Across the 926 patients with both preoperative and postoperative scores, the difference in the change scores between the full-length OSS and CAT was 0.21 points (mean) and 0.30 points (median).

Oxford Elbow Score. The CAT algorithm for the OES reduced the total length of the instrument from 12 items to a mean of 4.87 items (SD 0.33). In 13% of cases, four items were used by the CAT algorithm, and in 87% of cases five items were required. The difference in mean score between the full-length (12-item) OES and the CAT version was 0.13 out of 48 points. Pearson's correlation coefficient and the ICC for the total scale were both 0.98.

For each subscale, the mean difference between full-length subscale score and CAT score was less than 0.1 out of 16 points. Correlations between full-length subscale scores and CAT subscale scores ranged from 0.95 to 0.98 (Table II).

Preoperative and postoperative scores were available for 308 patients. The mean difference in change scores between full-length OES and CAT was 1.58 points (SD 2.71) (median 1.35 points (IQR -0.35 to 3.06)).

We have provided kernel density plots, scatter plots, and Bland-Altman plots for all CAT comparisons as Supplementary Material.

Discussion

We have developed CATs for the suite of Oxford scores, specifically the OHS, OKS, OSS, OES, and OES subscales. These are among the most widely used PROMs in their fields, especially in the UK and Europe,[10,20] and are commonly used in areas where the NIHR has recognized strains that may contribute to PROMs not gaining the traction they need to improve patient care. We have markedly reduced the assessment lengths of the Oxford scores, in many instances by half. In pressurized clinical systems, this may translate into saving time, and meaningful increases in uptake and completion. In research studies, several outcome measures are often used, and reducing the burden of some or all of them might contribute to gaining more complete study data. At the same time, this advantage is achieved without compromising on score accuracy, as assessed in the battery of tests that we used.

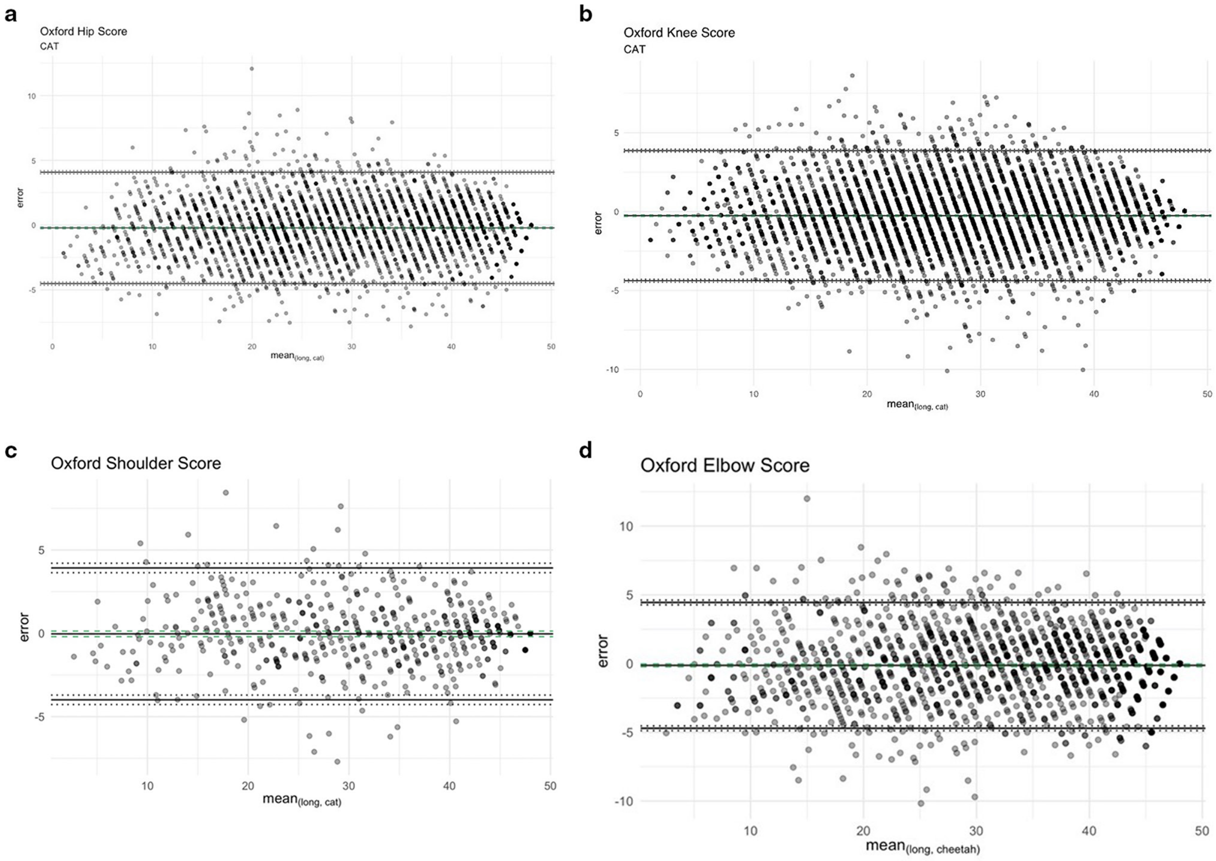

Fig. 3

a) Bland-Altman plot demonstrating the concordance of scores from the full-length and computerized adaptive testing (CAT) versions of the Oxford Hip Score. For each pair of scores, the x-axis represents the mean of both scores and the y-axis represents the difference between them. The outermost, horizontal solid lines represent the 95% limits of agreement and the surrounding dotted lines represent their 95% confidence intervals. b) Bland-Altman plot demonstrating the concordance of scores from the full-length and CAT versions of the Oxford Knee Score. c) Bland-Altman plot demonstrating the concordance of scores from the full-length and CAT versions of the Oxford Shoulder Score. d) Bland-Altman plot demonstrating the concordance of scores from the full-length and CAT versions of the Oxford Elbow Score (total combined score of all three subscales).

To provide context for our results, the ICC between the OBERD CAT and the full-length questionnaires was between 0.95 and 0.98, whereas the original development study of OKS demonstrated a test-retest Pearson's correlation coefficient of 0.92.[5] This suggests that our CATs perform well, relative to the intrinsic variability in PROM scores seen when a person completes the questionnaire twice. Furthermore, the difference between OBERD CAT score estimations and full-length questionnaire scores are unlikely to be important in the real world. This good CAT performance was demonstrated across the range of potential scores, as confirmed by plotting the score distributions for each PROM (a selection of illustrative figures has been included in this paper). We compared the difference between our CAT scores and the full-length questionnaire with the MCID, and the differences between OBERD CATs and full-length scores were generally smaller than MCIDs for the PROMs concerned. While the accuracy of CAT assessments was high at the population-level, and generally high at the individual level, there will be some cases where an individual's CAT score and full-length assessment score differ by more than the MCID. The impact of these unusual cases may be greater for studies with small sample sizes.

In some instances, CAT-based reduction in items posed was not dramatic. Nevertheless, contemporary electronic PROM administration still offers advantages over paper-based administration. Our system can be used on computers, tablets, smartphones, and with automated voice communication. Even if major item reduction is not achieved for a proportion of people, these features may still be appealing. The ability to complete CATs on the individual's personal smartphone, rather than require healthcare IT infrastructure, may also be advantageous. When differences arose between full-length questionnaire and CAT, the cases exhibiting the largest discrepancies were the most atypical in

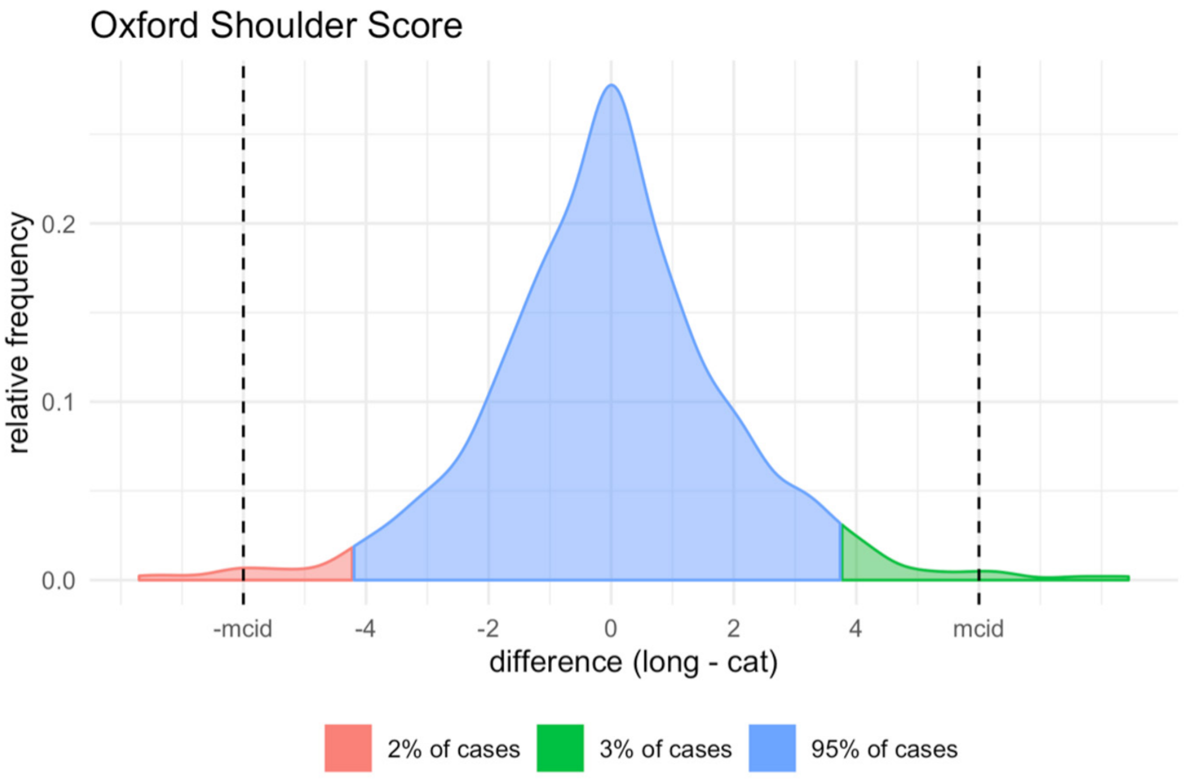

Fig. 4

Difference in score between the full-length Oxford Shoulder Score and its computerized adaptive testing (CAT) counterpart. Dashed horizontal lines represent the instrument's minimal clinically important difference (MCID).

the training set, because there are not enough similar cases to achieve better accuracy by machine-learning methods.

There are other outcome measures in these clinical areas,[10,20] including some PROM options that share CAT systems advantages, such as the Patient-Reported Outcomes Measurement Information System (PROMIS), which employs item response theory.[21] Item response theory-based CATs involve referencing an individual against a dataset in real time. In practice, this is acknowledged to lead to time lag for the user completing the CAT.[22] The regression tree approach that we used to develop OBERD CATs is known to avoid this.[22] This could be formally assessed in a future comparative study, though real-world deployment and user feedback has not revealed performance issues to date.

Furthermore, our CAT algorithms use the existing items of well-accepted PROMs. The Oxford scores that we have studied here are widely used in the UK and Europe.[10,20] Oxford scores are also used outside Europe, for example in New Zealand's arthroplasty registry.[10] They are generally well-developed, validated, and familiar to clinicians. Furthermore, we are proposing the use of our CATs in scenarios where these PROMs are already used.

There are strengths to our study design. In particular, we used large datasets, and split the data into completely separate training and test datasets. As a result, our final analyses are at less risk of instability and imprecision, and provide an exploration of the generalizability of our findings.[23] Similar methodological approaches have been used to successfully develop CATs for other musculoskeletal PROMs.[24,25] Such successes do not guarantee that good performance will be seen with other measures, hence the value of our study in applying these techniques to the Oxford scores.

There are also limitations to our study. We used pre-existing data and these may not be generalizable to other groups, who differ from our cohort based on demographics, diagnoses, or treatments. To account for overfitting, we employed a training/test data split, in which data leakage was prevented to ensure the results are based on applying the algorithms to fresh data. We undertook randomization of forms before establishing whether forms were complete. This led to exclusion of some forms after randomization. This sequence could be reversed in future, though we do not believe it introduced bias into the results.

The CAT is currently implemented as a web-based application, so is available to any device with an internet connection and a browser. Future work might focus on examining the real-world impact of CATs like these. While CATs harness questions that will provide the most information, qualitative research might explore whether the content validity of CAT assessments is perceived to be comprehensive and relevant. Additionally, there remains uncertainty around whether reducing the burden of assessment will translate into better compliance in terms of patients completing postoperative surveys. However, this is plausible, and the benefits of CATs are acknowledged by bodies such as the American Academy of Orthopaedic Surgeons.[26] Furthermore, the acceptability of our CATs could be tested in routine practice and research. This could be achieved by deploying some or all the OBERD CATs in routine clinical practice in a national registry or similar. As with all CATs, OBERD CATs are completed electronically, rather than with pen and paper. Traditionally, this has been considered to limit access for those who do not use technology and might affect acceptability. However, this is becoming rarer. In the UK, smartphone access has increased year on year, to as high as 92% in 2021.[27] While smartphone ownership may be distinct from digital literacy, we believe that the level of digital literacy likely to be needed to complete an electronic PROM is low. Furthermore, there is no difference in the level of digital literacy needed to complete our CATs compared to completing full-length electronic PROMs. As a result, we believe it is increasingly unlikely to be an issue affecting the rollout of systems like this, which have potential in both clinical practice and research.

In summary, we have developed machine-learning-based CATs for the OHS, OKS, OSS, and the OES, and its subscales. These CATs all reduced the burden of items to be completed, without compromising on score accuracy. Moreover, the validity of the questions themselves is supported by the success of the original instruments. Given this, and potential efficiencies of machine learning CATs, we believe that OBERD CATs will prove a useful option for measuring musculoskeletal outcomes in clinical practice and research.

Take home message

- Computerized adaptive tests (CATs) of the Oxford Hip, Knee, Shoulder, and Elbow Scores can markedly reduce the burden of completing these patient-reported outcome measures.
- CATs of Oxford scores do not sacrifice accuracy of measurement.

Twitter

Follow C. J. Harrison @conrad_harrison
Follow J. N. Rodrigues @mrjnrodrigues

Supplementary material

Kernel density plots, scatterplots, and Bland-Altman plots comparing full-length scores and computerized adaptive testing scores for each measure.

References

1. **No authors listed**. Patient reported outcome measures (PROMs). NHS Digital. https://www.england.nhs.uk/statistics/statistical-work-areas/proms/ (date last accessed 22 August 2022).
2. **Calvert M**, **Kyte D**, **Mercieca-Bebber R**, **et al**. Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO Extension. *JAMA*. 2018;319(5):483–494.
3. **Ishaque S**, **Karnon J**, **Chen G**, **Nair R**, **Salter AB**. A systematic review of randomised controlled trials evaluating the use of patient-reported outcome measures (PROMs). *Qual Life Res*. 2019;28(3):567–592.
4. **Denis F**, **Basch E**, **Septans A-L**, **et al**. Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. *JAMA*. 2019;321(3):306–307.
5. **Dawson J**, **Fitzpatrick R**, **Murray D**, **Carr A**. Questionnaire on the perceptions of patients about total knee replacement. *J Bone Joint Surg Br*. 1998;80-B(1):63–69.
6. **Murray DW**, **Fitzpatrick R**, **Rogers K**, **et al**. The use of the Oxford hip and knee scores. *J Bone Joint Surg Br*. 2007;89(8):1010–1014.
7. **Dawson J**, **Fitzpatrick R**, **Carr A**, **Murray D**. Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg Br*. 1996;78-B(2):185–190.
8. **Dawson J**, **Rogers K**, **Fitzpatrick R**, **Carr A**. The Oxford shoulder score revisited. *Arch Orthop Trauma Surg*. 2009;129(1):119–123.
9. **Dawson J**, **Doll H**, **Boller I**, **et al**. The development and validation of a patient-reported questionnaire to assess outcomes of elbow surgery. *J Bone Joint Surg Br*. 2008;90-B(4):466–473.
10. **Rolfson O**, **Bohm E**, **Franklin P**, **et al**. Patient-reported outcome measures in arthroplasty registries Report of the Patient-Reported Outcome Measures Working Group of the International Society of Arthroplasty Registries Part II. Recommendations for selection, administration, and analysis. *Acta Orthop*. 2016;87 Suppl 1:9–23.
11. **Greenhalgh J**, **Long AF**, **Flynn R**. The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory? *Soc Sci Med*. 2005;60(4):833–843.
12. **Greenhalgh J**, **Dalkin S**, **Gooding K**, **et al**. Functionality and feedback: a realist synthesis of the collation, interpretation and utilisation of patient-reported outcome measures data to improve patient care. *Health Serv Deliv Res*. 2011;5(2):1–280.
13. **Loh W**. Classification and regression trees. *WIREs Data Mining Knowl Discov*. 2011;1(1):14–23.
14. **Bland JM**, **Altman DG**. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135–160.
15. **Plummer OR**, **Abboud JA**, **Bell J-E**, **et al**. A concise shoulder outcome measure: application of computerized adaptive testing to the American Shoulder and Elbow Surgeons Shoulder Assessment. *J Shoulder Elbow Surg*. 2019;28(7):1273–1280.
16. **Beaton DE**, **Boers M**, **Wells GA**. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol*. 2002;14(2):109–114.
17. **Beard DJ**, **Harris K**, **Dawson J**, **et al**. Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. *J Clin Epidemiol*. 2015;68(1):73–79.
18. **Dawson J**, **Doll H**, **Boller I**, **et al**. Comparative responsiveness and minimal change for the Oxford Elbow Score following surgery. *Qual Life Res*. 2008;17(10):1257–1267.
19. **Jones IA**, **Togashi R**, **Heckmann N**, **Vangsness CT**. Minimal clinically important difference (MCID) for patient-reported shoulder outcomes. *J Shoulder Elbow Surg*. 2020;29(7):1484–1492.
20. **Hawkins RJ**, **Thigpen CA**. Selection, implementation, and interpretation of patient-centered shoulder and elbow outcomes. *J Shoulder Elbow Surg*. 2018;27(2):357–362.
21. **No authors listed**. PROMIS. Health Measures. https://www.healthmeasures.net/explore-measurement-systems/promis (date last accessed 22 August 2022).
22. **Delgado-Gómez D**, **Laria JC**, **Ruiz-Hernández D**. Computerized adaptive test and decision trees: a unifying approach. *Expert Systems with Applications*. 2019;117(1):358–366.
23. **James G**, **Witten D**, **Hastie T**, **Tibshirani R**. *An Introduction to Statistical Learning*. Springer, 2013: 426.
24. **Kane LT**, **Namdari S**, **Plummer OR**, **Beredjiklian P**, **Vaccaro A**, **Abboud JA**. Use of computerized adaptive testing to develop more concise patient-reported outcome measures. *JB JS Open Access*. 2020;5(1):e0052.
25. **Lee D**, **Rao S**, **Campbell RE**, **et al**. The application of computerized adaptive testing to the International Knee Documentation Committee Subjective Knee Evaluation Form. *Am J Sports Med*. 2021;49(9):2426–2431.
26. **Grogan Moore M**, **Jayakumar P**, **Koenig K**. AAOS now: not just for research anymore: the usefulness of PROMs in clinical practice. American Academy of

Orthopaedic Surgeons. https://www.aaos.org/aaosnow/2019/sep/managing/managing01/ (date last accessed 22 August 2022).

27. **O'Dea S**. UK: smartphone ownership by age from 2012-2021. https://www.statista.com/statistics/271851/smartphone-owners-in-the-united-kingdom-uk-by-age/#statisticContainer (date last accessed 8 August 2021).

Author information:
- C. J. Harrison, MRCS, Director
- J. N. Rodrigues, PhD, Director
  Methodology Oxford Limited, London, UK.
- O. R. Plummer, PhD, Chief Scientific Officer
- A. Hunt, BSc, Data Analyst
  Universal Research Solutions, Columbia, Missouri, USA.
- J. Dawson, DPhil, Visiting Researcher
- C. Jenkinson, DPhil, Professor of Health Services Research and Director of the Health Services Research Unit
  Nuffield Department of Population Health, University of Oxford, Oxford, UK.

Author contributions:
- C. Harrison: Visualization, Writing – original draft, Writing – review & editing.
- O. R. Plummer: Conceptualization, Formal analysis, Investigation, Supervision, Writing – review & editing.
- J. Dawson: Writing – review & editing.
- C. Jenkinson: Writing – review & editing.
- A. Hunt: Data curation, Formal analysis, Investigation, Writing –review & editing.
- J. N. Rodrigues: Visualization, Writing – original draft, Writing – review & editing.

Funding statement:
- The authors disclose receipt of the following financial or material support for the research, authorship, and/or publication of this article: funding provided by Universal Research Solutions, LLC.

ICMJE COI statement:
- C. Harrison and J. N. Rodrigues are Directors of Methodology Oxford Ltd, which was commissioned to assist with this work through Oxford University Innovation consultancy services. J. N. Rodrigues also reports project grants from NIHR (unrelated to this study), and is Chair of the British Society for Surgery of the Hand UK Hand Registry and an unpaid BSSH Council member. J. Dawson reports a grant from NIHR (Research for Patient Benefit ref PB-PG-1217-20017 1/10/19-30/09/21) for development and validation of the GCA PRO questionnaire, and royalties and consulting fees from Oxford University Innovation, all of which are unrelated to this study. C. Harrison is also funded by an NIHR Doctoral Research Fellowship (NIHR300684) for similar work, unrelated to this study.

Open access funding
- The open access fee for this study was funded by Oxford University Innovation.