Categorization of changes in the Oxford Knee Score after Total Knee Replacement: An interpretive tool developed from a dataset of 46,094 replacements.

Authors:

Mette Mikkelsen: Dept. of Orthopaedic Surgery, Clinical Orthopaedic Research Hvidovre (CORH), Copenhagen University Hospital Hvidovre, Kettegård Alle 30, 2650 Hvidovre, Copenhagen, Denmark.

CORRESPONDING AUTHOR
e-mail: mette.mikkelsen.02@regionh.dk
phone: +45 60172731

Anqi Gao: Nuffield Dept. of Orthopaedics, Rheumatology and Musculoskeletal Science, University of Oxford, Windmill Road, Headington, Oxford OX3 7LD, England.
e-mail: gangee2014@gmail.com

Lina Holm Ingelsrud: Dept. of Orthopaedic Surgery, Clinical Orthopaedic Research Hvidovre (CORH), Copenhagen University Hospital Hvidovre, Kettegård Alle 30, 2650 Hvidovre, Copenhagen, Denmark.
e-mail: lina.holm.ingelsrud@regionh.dk

David Beard: Nuffield Dept. of Orthopaedics, Rheumatology and Musculoskeletal Science, University of Oxford, Windmill Road, Headington, Oxford OX3 7LD, England.
e-mail: david.beard@ndorms.ox.ac.uk

Anders Troelsen: Dept. of Orthopaedic Surgery, Clinical Orthopaedic Research Hvidovre (CORH), Copenhagen University Hospital Hvidovre, Kettegård Alle 30, 2650 Hvidovre, Copenhagen, Denmark.
e-mail: anders.troelsen@regionh.dk

Andrew Price: Nuffield Dept. of Orthopaedics, Rheumatology and Musculoskeletal Science, University of Oxford, Windmill Road, Headington, Oxford OX3 7LD, England.
e-mail: andrew.price@ndorms.ox.co.uk

Abstract.

Objective To create an interpretive categorical classification for the transition in Oxford Knee Scores (OKS) change score (ΔOKS) using the anchor-based method.

Study design and setting Registry data from 46 094 total knee replacements from the year 2014/15, accessed via the Health and Social Care Information Centre (HSCIC) official website. Data included pre-operative and 6-month follow-up OKS and response to the transition anchor question. Categories were determined using Gaussian approximation probability and k-fold cross-validation.

Results 4 categories were identified with the corresponding ΔOKS intervals; "*1. Much Better*" (≥ 16), "*2. A Little Better*" (7-15), "*3. About the Same*" (1-6) and "*4. Much Worse*" (≤ 0) based on the anchor questions' original 5 categories. The mean 10-fold cross-validation error was 0.35 OKS points (95 % confidence interval 0.12 to 0.63). Sensitivity ranged from 0.34 to 0.68, specificity ranged from 0.74 to 0.95.

Conclusion We have categorized the change score into a clinically meaningful classification. We argue it should be an addition to the continuous OKS outcome to contextualize the results in a way more applicable to the shared decision making process and for interpreting research results.

Keywords

Oxford Knee Score; Patient reported outcome; Knee Replacement; Interpretive tool

Word count: 185

Running Title: Categorization of changes in the Oxford Knee Score

What is new?

- This study is to our knowledge the first to categorize the change in Oxford Knee Score (OKS).
- Using gaussian approximation probability is a new take on the methodology in this field, and adds to the discussion of a "best practice".
- This categorized change in OKS will enable an easier interpretation of research results, aiding communication with patients and non-healthcare professionals.

1. Introduction

The Oxford Knee Score (OKS) is one of the most common knee specific patient reported outcome measures (PROM) used to evaluate the outcome of knee replacement[1]. The continuous score offers rich, high granularity information, however, it can be harder to interpret and communicate to non-healthcare professionals than a categorical outcome. In 2016 Rolfson et al.[2] found that a total of 13 knee arthroplasty registries collected PROM data, either on all patients or on a sample population. Of these three registries had OKS as one of the PROMs, all of which included a 6 months follow up data collection[2]. As PROMs are increasingly included in national registries and randomized clinical trials there is also an increasing need for better interpretive tools[3–5]. Improved interpretation helps patient decision making and increases research finding validity. Whilst acknowledging the need, categorizing PROMs is a field where there is no consensus on "best practice" or how we insure they are applied correctly[6].

Categorization requires the setting of boundaries or "cut offs" and there are several ways to establish these boundary scores. Cut-offs for Patient Acceptable Symptom State (PASS) and Minimal Important Change (MIC) are usually determined using the anchor based method, which seems to be more widely accepted than the distribution based method. A number of different methods have been applied to statistically determine the boundary value, of which the Receiver Operating Characteristic (ROC) analysis is widely used [7,8]. We propose Gaussian approximation probability as an alternative. We argue Gaussian approximation is better for data in which there is a large effect size to counteract an uneven distribution of patients in the groups when determined by the anchor question. The ROC is based on a confusion matrix which makes it vulnerable to prevalence bias, with skewing towards the groups with a large number of observations when using the "best fit" approach[9,10]. In contrast, the Gaussian approximation uses the data to fit functions which are then used to calculate the cut-off between groups, thus minimizing prevalence bias. Classifications like MIC, PASS and the categories originally developed on the Oxford Hip Score proposed by Kalairajah et al.[11] are all widely used. Both PASS and Kalairajah's categories pertain to the final score, and the MIC interprets change. Kalairajah's categories divide the score into four categories; poor, fair, good and excellent based on cumulative frequency distribution. PASS introduces a measure of the treatments success, whereas the MIC concept aims to

determine the change score value where patients have an important clinical change (both benefit and detriment)[12]. An important concept for examining outcome scores is the distinction between final outcome (an absolute score) or how much the score has changed over a set time period (transition or delta change). To evaluate the outcome from knee replacement fairly, both transition and final outcome should be considered. We believe there is a need for a classification which interprets the transition in a meaningful way to patients and other non-healthcare professionals. We thus sought to expand on the MIC concept and define categories which capture the distribution of changes on a group level after knee replacement.

Our aim was to categorize the change in OKS from pre-operative to 6 months follow-up (ΔOKS) for interpretation of the transition in a way which describes the range of different outcomes available after knee replacement surgery.

2. Method

2.1 Data Source and Participants

The data set consisted of 46 094 total knee replacements from the April 1. 2014 to March 31. 2015, accessed via the National Health Service (NHS)-digital website. PROMs data has been collected by the Health and Social Care Information Centre (HSCIC) on all NHS funded knee replacements in England as part of the NHS PROMs program since 2009, mandated by the Department of Health. The PROMs are collected at baseline and at 6 months follow-up. 6-months was chosen because it was believed to be the earliest time point where an average patient had reached the clinically important benefits[13]. A total of 83 450 total knee replacements were done during that time, 47 743 had record level data available on the NHS-Digital website, and of these 46 094 (55%) had complete PROMs data and were included in the study. This data has previously been presented in A. Gao's thesis[14].

2.2 Study design

Categories were determined using the anchor-based method [15].Our variables were pre-operative and 6 months follow-up of both the OKS and patients' response to the transition anchor question:

1. "How are the problems now in the operated knee compared to pre-operation?" Please select one of the following: "much better", "a little better", "about the same", "a little worse" and "much worse".

The OKS is a 12 item knee specific PROM evaluating two domains; knee pain and function. Each question has 5 response options and the score range is 0-48 with high scores indicating low disability[16].

2.3 Statistics

To evaluate the assumption of an ordered distribution within the transition question's categories distribution of the change score in relation to the transition question was investigated through qualitatively assessing histograms, box plots and quartiles for each group. If groups overlapped to a degree where they were deemed interchangeable by the author group, they were collapsed into one group, to avoid violating the ordered assumption.

Thresholds between groups were determined using Gaussian approximation probability. This meant fitting normal approximation functions to each group. The thresholds were then determined by identifying the root between adjacent group's functions. Roots were identified using the *uniroot* function in R[17], which applies Brent's Method[18].

To validate the thresholds a 10-fold cross-validation was performed, as described by Casella et al.[19]. The 10-fold cross-validation randomly splits the dataset into ten smaller datasets and tests them individually against the remainder of the original dataset. It thereby assesses the analyses' ability to be generalized onto an independent cohort, by evaluating the accuracy of the estimates. This evaluation yields a cross validation error. We presented the cross validation error as a mean distance in OKS points from the test data prediction to that of the training data defined by the following equation:

$$cv.error = \frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{\left(y_i - \hat{f}(x_i)\right)^2}\right),$$

where y_i is the i^{th} training threshold and $\hat{f}(x_i)$ is the prediction for $\hat{f}$ made from the i^{th} smaller test set (fold). The thresholds chosen as final outcome were the mean over the 10 training data sets. 95 % confidence intervals (CI 95 %) were also calculated using the 10 training sets. A further measure of supervision of the anchor-based method we calculated sensitivity and specificity based on the one group against the rest approach.

The robustness of this method was evaluated by calculating the thresholds and CI 95 % for other established methods (ROC (Youden best fit), predictive modelling and adjusted predictive modelling) and compare the CI 95 % ranges, the narrower the range the more robust the method. The CI 95 % were calculated using 1000 bootstrap replications[10,20]. To assess the impact of the correlation between baseline OKS and the thresholds, we did sub-analysis stratifying patient into two groups based on the mean baseline OKS.

All statistics were calculated using R version 3.6.0[17].

3. Results

3.1 Participants:

46,094 patients were included in the study, 54.3 % female, 84.3 % were 60 years of age or older when they had their knee replacement and 77.1 % suffered from arthritis. Additional patient demographics are available in Table 1.

Variable	
	Number (%)
N	46094
Gender	
NA	2939 (6.4)
Male	18123 (39.3)
Female	25032 (54.3)
Age band	
NA	2939 (6.4)
40 – 49	62 (0.1)
50 – 59	4239 (9.2)
60 – 69	15835 (34.4)
70 – 79	17763 (38.5)
80 – 89	5256 (11.4)
	Mean (SD)
Pre-OP Oxford Knee Score	19.17 (7.74)
Pre-OP EQ-5D Index	0.42 (0.31)
Pre-OP EQ-5D VAS	68.42 (19.77)
	Number (%)
Comorbidity	
Arthritis	35519 (77.1)
Heart Disease	4573 (9.9)
Hypertension	20740 (45.0)
Stroke	778 (1.7)
Circulatory Disease	2969 (6.4)
Lung Disease	4276 (9.3)
Diabetes	5975 (13.0)
Kidney Disease	869 (1.9)
Nervous Disorder	466 (1.0)
Liver Disease	254 (0.6)
Cancer	2290 (5.0)
Depression	4068 (8.8)

Table 1: Patient demographics at time of Surgery
EQ-5D = EuroQol-5 dimensions questionnaire

3.2 Descriptive statistics to determine number of categories and appropriateness of statistical approach.

The distributions of OKS in the groups were ordered (Figure 1). "*3. About the same*" and "*4. A little worse*" were collapsed into one category labelled "*3. About the same*" due to very similar distributions, ranges and because both categories represent minimal or no improvement but also do not need to be identified as having a substantially inferior outcome.

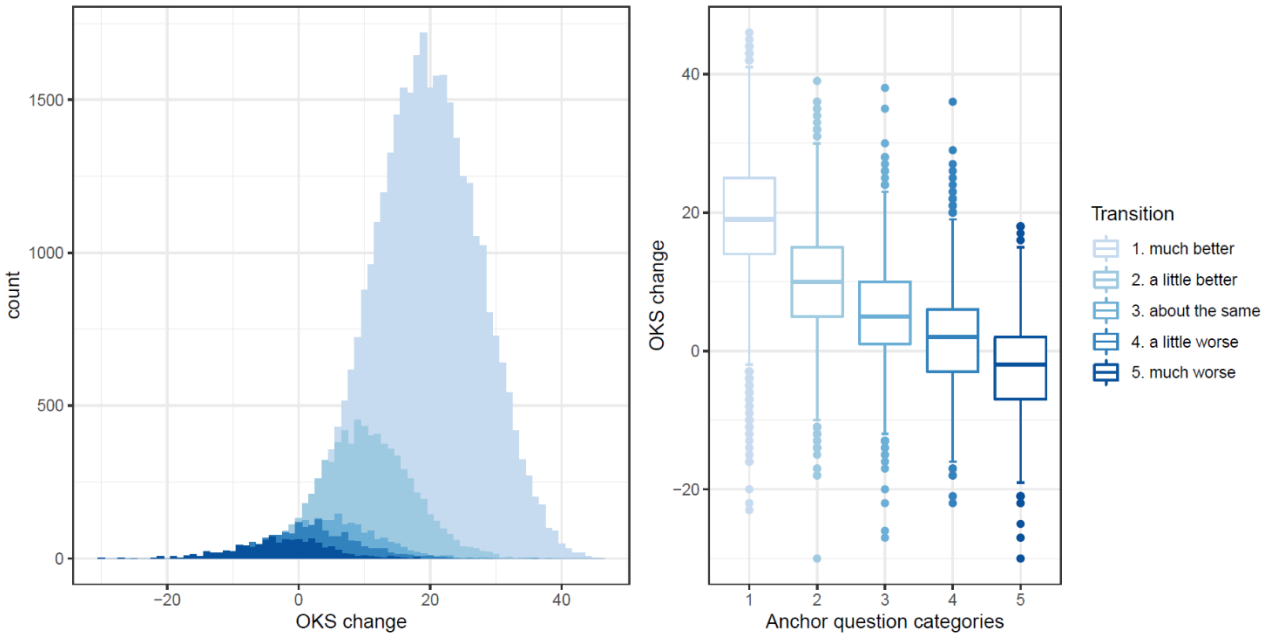
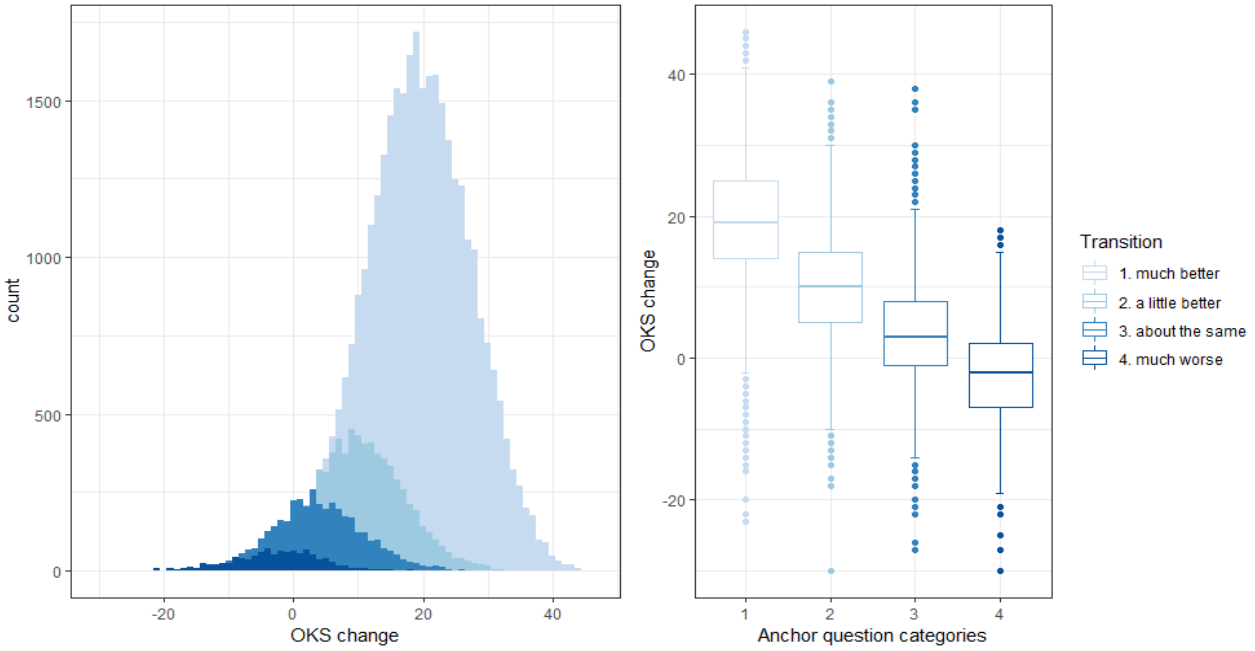


Figure 1. The distribution of change scores within the anchor question categories. OKS = Oxford Knee Score.

Figure 2 and Table 2 show the distributions of the final four groups. All groups displayed an approximate normal distribution (Figure 2) and the smallest group contained 1050 patients (Table 2), thus the methods assumptions were satisfied.



153

154 **Figure 2. The distribution within the categories after groups "about the same" and "a little worse" have been combined. OKS =**
155 **Oxford Knee Score**

156

	Mean	1st Quartile	Medial	4th Quartile	N (%)
1. Much better	19.4	14	19	25	33637 (73.0)
2. A little better	10.02	5	10	15	7521 (16.3)
3. About the same	3.78	-1	3	8	3886 (8.4)
4. Much worse	-2.69	-7	-2	2	1050 (2.3)

157 **Table 2: Distribution of change in Oxford Knee Score across the anchor questions' 4 final groups.**

158

	Change Score Interval
1. Much better	≥ 16
2. A little better	7 – 15
3. About the same	1 – 6
4. Much worse	≤ 0

159 **Table 3: Oxford Knee Scores' change score intervals derived from the calculated**
160 **thresholds.**
161

162

3.3 The classification and its thresholds.

The classifications' final four categories are presented in table 3 and ranges from "much worse" to "much better". We identified the cut-offs and their CI 95 % using the 10-fold cross-validation. ΔOKS = 15.11 (CI 95 % 15.07-15.14) marked the cut-off between "*1. Much better*" and "*2. A little better*". Between "*2. A little better*" and "*3. About the same*" it was ΔOKS = 6.98 (CI 95 % 6.91-7.02). And between "*3. About the same*" and "*4. Much worse*" it was ΔOKS = 0.56 (CI 95 % 0.47-0.62). The cross-validation error presented as mean distance from test to training data in OKS points was 0.35 (CI 95 % 0.12 to 0.63), visualized in Figure 3. The sensitivity ranged from 0.34 to 0.68, and the specificity ranged from 0.74 to 0.95 (Table 4).

The robustness of our method compared to established methods is presented in Table 1 in Supplementary Materials. It showed Gaussian approximation has narrower CI 95 % than the other methods for all thresholds.

The sub-group analysis based on baseline OKS showed a clinical significant difference, based on the minimal important difference (MID) of 5 points, between the groups. With patients demanding a larger change if they have a low baseline OKS (Table 2 in Supplementary Materials).

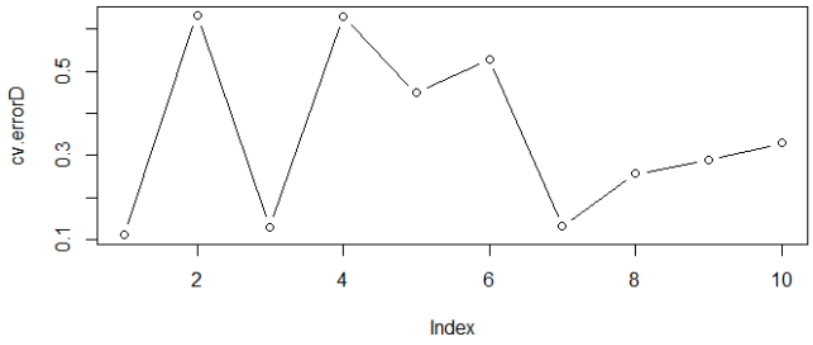


Figure 3. Cross validation error (cv.errorD) as distance in OKS for all 10 test folds (index).

	Sensitivity	Specificity
1. Much better	0.68	0.85
2. A little better	0.47	0.74
3. About the same	0.34	0.92
4. Much worse	0.64	0.95

Table 4: Sensitivity and specificity of the change score intervals.

4. Discussion

We set out to categorize the change score, using a 5 category anchor questions, into a tool for interpretation of research results. We developed the 4 category classification of the transition in OKS presented in table 3, describing the range of outcomes possible for knee replacement surgery, from “much worse” to “much better”. Improvements 16 points or above represents a change described as “much better” by the patients. The relatively low threshold of 16 points allows patients with a higher pre-operative score to still achieve a large perceived benefit from the intervention. However a baseline score of 33 makes it impossible to achieve a “much better” transition using this classification. This is partly the reason we argue both the final outcome and the transition should be evaluated.

The score distributions within the categories were ordered. "*3. About the same*" and "*4. A little worse*" were collapsed into one group; “*3. About the same*” with the range ΔOKS 1-6. As a result 4 categories were identified (table 3).The intervals displayed large variation in range size, meaning the change between groups is not linear. The non-linearity cause the relative elevation in change score from “about the same” to “a little better” to be small, but a corresponding change to “much better” requires more than double the change score than to be described as “a little better”. This non-linearity also adds to explanation for the low sensitivity in the “about the same” to “a little better” groups.

There are two methods for categorizing PROMs, the anchor-based method and the distribution-based method. The anchor-based method displays a large number of strengths compared to the distribution-based method. It has potentially greater clinical relevance as the categorization is referenced to the patients’ own experience. As a result the resultant categorization is more relatable and easy to communicate to patients. In effect this reduces the risk of assigning clinical meaning to a statistical term, and is contrary to the distribution method in this context. Consequently the anchor-based method has become the dominant approach for other PROM classifications e.g. PASS and MIC.

The transition in OKS score was evaluated at 6 months. To some extent this was a forced time point for the analysis and was the time used by the NHS PROMs program based on clinical rationale. Six months was considered sufficiently far enough into the recovery window that patients could adequately report on near final outcome, but still close enough to the time of surgery to prevent substantial recall bias and contamination[13]. In recent years a case for using 1 year rather than 6 months has been made[2,13]. Furthermore, studies indicate a time dependency for MIC and PASS[22,23], thus these tools should only be used for PROMs assessed

at the same time point for which they were designed for. It could be beneficial to calculate values for the 1 year follow-up if available, but this will be dependant on how it aligns with the collection for various national registries[2,3,5].

Our data is from 2014-2015, and is now 5 years old. A point about obsolescence could be argued but we feel it has little consequence for the validity of these categories since there have been no major changes to design, practice or outcome for TKA in the last 5 years[21].

When assessing the performance of the categories in terms of sensitivity and specificity, we largely avoid misclassification in the "*1. Much better*" and "*4. Much worse*" groups. However, "*2. A little better*" and "*3. About the same*" have larger misclassification errors. These two groups have the smallest score ranges and have adjacent groups in both ends of their ranges, meaning their misclassifications are not dichotomous. This complicates the interpretation of the misclassification, and means the patients can either be falsely classified to 1) having a more positive outcome or 2) a more negative outcome. Using the "*3. About the same*" group as an example, the high specificity means that 92% of those who did were in the group by using the criteria score range 1 – 6 do consider themselves to be "about the same". In other words: 8% of those who did not find themselves to be "about the same" where incorrectly classified as such. The sensitivity is much lower, indicating there are quite a large proportion of patients (66%) who in fact did consider to be "about the same", but when using the criteria OKS change 1-6, they are misclassified into another category.

The variety in sensitivity observed is largely due to the overlap between groups seen in figure 2, and their small ranges. A way to increase the sensitivity could be to combine the "*2. A little better*" and "*3. About the same*" groups, reducing the misclassification by increasing the ranges and decreasing the number of groups. We have chosen not to because we believe the cut between these two groups describes a clinically important threshold the MIC. What gives us confidence in this threshold is its similarity to the MIC calculated using the same anchor in a previous publication (OKS = 6.5). And its similarity to the only other published MIC using the anchor-based method (OKS = 8), but at 1 year follow-up with a different anchor-question and a cohort from a different country [15,20]. Clement et al. found that the average patient improved 1.1 points from 6 months to 1 year follow-up, which makes an increase in MIC of approximately 1 point likely. An additional MIC value of 3.8 points at 6 months follow up was calculated by Browne et al, who used the mean change method and calculated the MIC as the difference in change scores between those responding "a little better" and "unchanged". The methodological dissimilarity explains their smaller MIC value. The sensitivity and specificity

corresponds to those found in studies which determined a MIC value using the same anchor question, which could indicate the misclassification also has to do with the anchor question itself[15]. Exploring this further, one reason the anchor might partially be to blame, is the five transition response structure. Using seven responses might have increased the sensitivity, not by making a classification with additional categories, but by reducing the impact of erroneously choosing outside the true classification by diluting the giving nuance to choose the right groups to combine. The argument for seven response categories is that the distance from "a little" to "much" is too large, however adding two more response options might give the opposite problem, where patients find it harder to distinguish between two categories. Anchors used for determining MIC have as few as three or as much as fifteen response options[12,24]. Lastly sensitivity and specificity are based on a confusion matrix making them sensitive to unbalanced numbers of observations in the groups[9], which was the reason we opted not to use the ROC approach.

The robust methodology and the large number of patients in this study are its largest strengths. Using Gaussian approximation is new in a PROMs context. It makes us less vulnerable to skewing of the thresholds compared to e.g. ROC analysis, due to its independence of the relative group sizes [9].We did 10-fold cross validation to determine the predictive robustness of it and found a cross-validation error of 0.35 OKS points. This indicates a threshold determined on a different dataset with different baseline characteristics would on average vary less than 0.35 points from the ones determined here, showing a high degree of validity. However before using this tool, one should compare the baseline characteristics to those from this cohort. Other methods for determining thresholds, using the anchor base method, are ROC and predictive modelling and adjusted predictive modelling. To further contextualize Gaussian approximations robustness we calculated the thresholds using all four methods, and used the range of the CI 95 % as the measure of robustness. Compared to all three alternatives and for all thresholds Gaussian approximation produced the most narrow CI 95 % (Table 1 in supplementary materials).

One limitation of the study methodology is that we cannot control for the potential effect of the baseline pre-operative scores on the change score, since the baseline is part of our outcome. This is seen in the sub-analysis based on this where thresholds are significantly smaller for patient with a higher than average baseline OKS (Table 2 in Supplementary Materials). This has two consequences; first, using this classification one should compare baseline scores with this cohort beforehand. And second, interpretation on an individual level is discouraged and should only be done on a group level. Building on this, the lack of BMI in this cohort limits the comparison of other cohorts to ours. Limitations connected to the OKS are that, patients might misinterpret

the score and answer it consistently with opposite directionality. This misinterpretation is likely part of the reason for the outliers in the "*1. Much better*" and "*4. Much worse*" groups in figure 2. Regarding the anchor question we are aware that recall bias and response shifts are present, and account for part of the misclassification seen in the sensitivity (table 4) [25]. As discussed earlier in this section the anchor itself can be a reason for misclassification. Other factors unknown to us are influencing the patient perceived treatment benefit and might also explain part of the misclassifications. However, the cross-validation error gives us an indication on how sensitive the thresholds are to changes in baseline characteristics. Furthermore, the large number of loss to follow-up (45%) makes this study vulnerable to attrition bias. To, investigate this we compared the age and gender distributions, to those published by the National Joint Registry for England, qualitatively we did not see a difference between our cohort and the complete cohort for the NJR including 2015 (Table 3 in supplementary materials).

In summary we have presented a categorization method which we believe has value in interpreting treatment outcome, presenting research results and evaluating clinical practice. We believe the strength of this approach is the ability to interpret transition at a group/population level, rather than on the patient specific level. On a group-level, the categorization is applicable to assign individual patients into the categorical treatment response, and can therefore be used for responder analysis. However, we still encourage conclusions being made on a group level, and not on a single individual.

In conclusion, we present a clinically meaningful four category classification of the change score, which is easy to apply to any data set and understandable for patients. We believe this classification can be a useful addition to the current continuous OKS outcome and can be used for contextualizing results to help the interpretation of research outcome.

Declaration of interest

One or more of the authors have received or will receive benefits for personal or professional use from a commercial party related indirectly to the subject of this article.

Funding Sources:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Harris K, Dawson J, Gibbons E, Lim CR, Beard DJ, Fitzpatrick R, et al. Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty 2016:101–8.

[2] Rolfson O, Eresian Chenok K, Bohm E, Lübbeke A, Denissen G, Dunn J, 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 I. Overview and rationale for patient-reported outcome measures. Acta Orthop 2016;87:3–8. doi:10.1080/17453674.2016.1181815.

[3] NJR. NJR's 15th Annual Report 2018;1821.

[4] Beard DJ, Davies LJ, Cook JA, MacLennan G, Price A, Kent S, et al. The clinical and cost-effectiveness of total versus partial knee replacement in patients with medial compartment osteoarthritis (TOPKAT): 5-year outcomes of a randomised controlled trial. Lancet 2019;6736:5–9. doi:10.1016/s0140-6736(19)31281-4.

[5] New Zealand orthopedic association. The New Zealand Joint Registry, 19th report; January 1999 to December 2017 2018.

[6] Price AJ, Alvand A, Troelsen A, Katz JN, Hooper G, Gray A, et al. Knee replacement. Lancet 2018;392:1672–82. doi:10.1016/S0140-6736(18)32344-4.

[7] Kvien TK, Heiberg T, Hagen KB. Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): What do these concepts mean? Ann Rheum Dis 2007;66:40–1. doi:10.1136/ard.2007.079798.

[8] Keurentjes JC, Van Tol FR, Fiocco M, So-Osman C, Onstenk R, M M Koopman-Van Gemert AW, et al. Patient acceptable symptom states after total hip or knee replacement at mid- term follow-up THRESHOLDS OF THE OXFORD HIP AND KNEE SCORES. Bone Jt Res 2014;33:7–13. doi:10.1302/2046-3758.31.2000141.

[9] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32–5.

[10] Terluin B, Eekhout I, Terwee CB. The anchor-based minimal important change, based on receiver operating characteristic analysis or predictive modeling, may need to be adjusted for the proportion of improved patients. J Clin Epidemiol 2017;83:90–100. doi:10.1016/j.jclinepi.2016.12.015.

[11] Kalairajah Y, Azurza K, Hulme C, Molloy S, Drabu KJ. Health outcome measures in the evaluation of total hip arthroplasties - A comparison between the harris hip score and the Oxford hip score. J Arthroplasty 2005;20:1037–41. doi:10.1016/j.arth.2005.04.017.

[12] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–15. doi:10.1016/0197-2456(89)90005-6.

[13] Browne JP, Bastaki H, Dawson J. What is the optimal time point to assess patient-reported recovery after hip and knee replacement? A systematic review and analysis of routinely reported outcome data from the English patient-reported outcome measures programme. Health Qual Life Outcomes 2013;11:1. doi:10.1186/1477-7525-11-128.

[14] Gao A, Beard D, Price A. Analysis of Patient-Reported Outcome Following Hip and Knee Replacement — Interpretation, Strengths, Limitations and Solutions. University of Oxford, 2019.

[15] Beard DJ, Harris K, Dawson J, Doll H, Murray DW, Carr AJ, et al. Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. J Clin Epidemiol 2015;68:73–9. doi:10.1016/j.jclinepi.2014.08.009.

[16] Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the Perceptions of Patients about Total Knee Replacement. J Bone Joint Surg Br 1998;80:63–9. doi:http://dx.doi.org/10.1302/0301-620X.80B1.7859.

[17] R Core Team. R: A language and environment for statistical computing. 2019.

[18] Brent RP. Algorithms for Minimization without Derivatives. Englewood Cliffs, N.J. : Prentice-Hall, [1973].; 1973.

[19] Casella G, Fienberg S, Olkin I. An Introduction to Statistical Learning. 2013. doi:10.1016/j.peva.2007.06.006.

[20] Ingelsrud LH, Roos EM, Terluin B, Gromov K, Husted H, Troelsen A. Minimal important change values for the Oxford Knee Score and the Forgotten Joint Score at 1 year after total knee replacement. Acta Orthop 2018;89:541–7. doi:10.1080/17453674.2018.1480739.

[21] Brittain R, Dawson-bowling S, Goldberg A, Toms A, Young E, Mccormack V, et al. NJR 17th Annual Report 2020.

[22] Galea VP, Rojanasopondist P, Connelly JW, Bragdon CR, Huddleston JI, Ingelsrud LH, et al. Changes in Patient Satisfaction Following Total Joint Arthroplasty. J Arthroplasty 2019. doi:10.1016/j.arth.2019.08.018.

[23] Galea V, Florissi I, Rojanasopondist P, Connelly JW, Ingelsrud LH, Bragdon C, et al. The Patient Acceptable Symptom State for the Harris Hip Score Following Total Hip Arthroplasty Validated thresholds at 3 months, 1, 3, 5, and 7 years follow-up. J Arthroplasty 2019. doi:10.1016/j.arth.2019.08.037.

[24] Rodrigues JN, Zhang W, Scammell BE, Davidson D, Fullilove S, Chakrabarti I, et al. Recovery, responsiveness and interpretability of patient-reported outcome measures after surgery for Dupuytren's disease. J Hand Surg Eur Vol 2017;42:301–9. doi:10.1177/1753193416677712.

[25] Norman G. Hi! How are you? Response shift, implicit theories and differing epistemologies. Qual Life Res 2003;12:239–49. doi:10.1023/A:1023211129926.

Supplementary materials:

Tabel 1 Comparison of thresholds calculated by four different methods.

	Gaussian (CI 95 %)	ROC_{best} (CI 95 %)	Predicted (CI 95 %)	Adjusted (CI 95 %)
"Much better" – "A little better"	15.11 (15.07-15.14)	15.5 (14.5-15.5)	14.64 (14.54-14.73)	14.01 (13.92-14.11)
"A little better" – "About the same"	6.98 (6.91-7.02)	6.5 (5.5-7.5)	6.90 (6.75-7.04)	6.63 (6.48-6.78)
"About the same" – "Much worse"	0.56 (0.47-0.62)	2.5 (-0.5-2.5)	0.57 (0.33-0.81)	0.02 (-0.20-0.25)

Tabel 2 Sub-group analysis based on baseline Oxford Knee Score, stratified by the mean score (19.17).

	Threshold	Threshold value (CI 95 %)	Cross-validation error (CI 95 %)
Low			0.36 (0.15-0.61)
	"Much better" – "A little better"	18.17 (18.13-18.20)	
	"A little better" – "About the same"	9.40 (9.32-9.46)	
	"About the same" – "Much worse"	3.21 (3.12-3.28)	
High			0.43 (0.20-0.72)
	"Much better" – "A little better"	10.78 (0.20-0.72)	
	"A little better" – "About the same"	3.34 (3.23-3.44)	
	"About the same" – "Much worse"	-3.76 (-3.87 - -3.66)	

Tabel 3 Comparison of patient demographics to assess attrition bias.

	PROMs cohort	NJR cohort
N	46094	772809
Syrgery year	2014-2015	2003-2015
Gender		
NA	6.4 %	0 %
Male	39.3 %	43 %
Female	54.3 %	57 %
Age median (IQR)		70 (64-76)
Age band		
NA	6.4 %	
40 – 49	0.1 %	
50 – 59	9.2 %	
60 – 69	34.4 %	
70 – 79	38.5 %	
80 – 89	11.4 %	