# Development of machine learning algorithms to predict clinically meaningful improvement for the patient-reported health state after total hip arthroplasty

Kunze, K.N.; Karhade, A.V.; Sadauskas, A.J.; Schwab, J.H.; Levine, B.R.

The Journal of Arthroplasty 35 (2020) 2119–2123

Contents lists available at ScienceDirect

The Journal of Arthroplasty

journal homepage: www.arthroplastyjournal.org

Primary Hip

# Development of Machine Learning Algorithms to Predict Clinically Meaningful Improvement for the Patient-Reported Health State After Total Hip Arthroplasty

Kyle N. Kunze, BS [a,*], Aditya V. Karhade, BE [b], Alex J. Sadauskas, MD [a], Joseph H. Schwab, MD, MS [b], Brett R. Levine, MD, MS [a]

[a] Department of Orthopaedic Surgery, Rush University Medical Center, Chicago, IL
[b] Department of Orthopedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA

ARTICLE INFO

*Article history:*
Received 4 February 2020
Received in revised form
1 March 2020
Accepted 10 March 2020
Available online 18 March 2020

*Keywords:*
total hip arthroplasty
THA
machine learning
clinical outcomes
MCID
clinically significant outcome

ABSTRACT

*Background:* Failure to achieve clinically significant outcome (CSO) improvement after total hip arthroplasty (THA) imposes a potential cost-to-risk imbalance in the context of bundle payment models. Patient perception of their health state is one component of such risk. The purpose of the current study is to develop machine learning algorithms to predict CSO for the patient-reported health state (PRHS) and build a clinical decision-making tool based on risk factors.

*Methods:* A retrospective review of primary THA patients between 2014 and 2017 was performed. Variables considered for prediction included demographics, medical history, preoperative PRHS, and modified Harris Hip Score. The minimal clinically important difference (MCID) for the PRHS was calculated using a distribution-based method. Five supervised machine learning algorithms were developed and assessed by discrimination, calibration, Brier score, and decision curve analysis.

*Results:* Of 616 patients, a total of 407 (69.2%) achieved the MCID for the PRHS. The random forest algorithm achieved the best performance in the independent testing set not used for algorithm development (c-statistic 0.97, calibration intercept −0.05, calibration slope 1.45, Brier score 0.054). The most important factors for achieving the MCID were preoperative PRHS, preoperative opioid use, age, and body mass index. Individual patient-level explanations were provided for the algorithm predictions and the algorithms were incorporated into an open access digital application available here: https://sorg-apps.shinyapps.io/THA_PRHS_mcid/.

*Conclusion:* The current study created a clinical decision-making tool based on partially modifiable risk factors for predicting CSO after THA. The tool demonstrates excellent discriminative capacity for identifying those at greatest risk for failing to achieve CSO in their current health state and may allow for preoperative health optimization.



The demand for total hip arthroplasty (THA) is estimated to grow by 174% to 572,000 annual procedures by the year 2030 [1]. Paralleling this exponential increase in THA procedures is a growing concern for unsustainable healthcare expenditures for the care regarding these procedures. Bundled payment care initiatives are now being employed for cost-control in anticipation of the overwhelming numbers and economic risk. To this end, increasing attention has been turned toward developing predictive models for identifying and optimizing patients who are at risk of readmission and incurring additional cost after surgery [2,3]. Further knowledge of the costs of care throughout the period of risk may offer an opportunity for cost containment [4,5].

Current methods of understanding and measuring patient improvement after THA are highly dependent on patient-reported outcome measures (PROMs). Administration and interpretation of scores generated by such PROMs may allow for preoperative identification of patients at risk for readmission or adverse outcomes based on the potential need for more complex care or

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.arth.2020.03.019.

* Reprint requests: Kyle N. Kunze, BS, Department of Orthopaedic Surgery, Rush University Medical Center, 1611 W. Harrison Street, Suite 300, Chicago, IL 60612.

https://doi.org/10.1016/j.arth.2020.03.019

anticipation of a poor outcome. Furthermore, psychometric-based applications of PROMs, which allow for the understanding of those who will experience a clinically meaningful improvement, may also help maximize the cost-effectiveness of bundle payment models. As evidence continues to be published for the use of PROMs within this context, it represents a feasible means of understanding patient risk, and further understanding could benefit orthopedic surgeons by minimizing cost concerns.

Preoperative mental health and postoperative fulfillment of expectations has known associations with patient outcomes after THA [6]. The patient-reported health state (PRHS), a visual analog scale derived from the EuroQol-5 dimension scale [7], is a commonly administered PROM which provides insight into a patient's perception of their current health state, and has been demonstrated to be associated with adverse outcomes [3]. Therefore, further understanding and application of this particular PROM could be of great benefit and allow for preoperative health optimization. Furthermore, to our knowledge, there are no preoperative clinical decision aids available for risk-stratifying patients based on this PROM.

The purpose of this study is to develop a clinical decision aid for patients undergoing primary THA to understand which patients will experience clinically significant improvement based on the PRHS after THA. The utility of supervised machine learning algorithms was investigated for this predictive analytic task, and the best performing algorithm was then deployed to allow for both prediction and an in-depth explanation of the quantified risk of not achieving clinically meaningful outcome at the individual patient level.

## Methods

### Data Source

Our Institutional Review Board approved retrospective review of electronic health records for this study. The inclusion criteria for the current study consisted of all patients undergoing primary THA for end-stage degenerative osteoarthritis of the hip between January 2014 and January 2016 (n = 731). Exclusion criteria included (1) etiology of hip osteoarthritis that was inflammatory, infectious, post-traumatic, or related to osteonecrosis (n = 45); (2) revision THA (n = 42); (3) patients with a history of neuromuscular disorder who could not participate in standard rehabilitation protocols, (4) failure to provide an answer to the PRHS (n = 28); and (5) less than minimum follow-up of 2 years. Data were subsequently collected from a total of 616 (84.3%) patients.

### Outcome

The primary outcome was minimal clinical important difference (MCID) for the PRHS at a minimum of 2 years postoperatively. The MCID represents minimal clinically significant change occurring beyond expected variance or error for a patient in the postoperative period for a specific outcome of interest. The MCID was defined using a distribution-based method to better represent the specific patient population in the current study [8], and this method has been shown to be reliable for a given study population widely throughout the orthopedic literature [9–12]. Briefly, the MCID threshold was calculated as 1/2 the standard deviation of postoperative PRHS score and applied to differences between the postoperative and preoperative PRHS to calculate the frequency of patients who achieved the MCID.

### Variables

Candidate variables were collected prospectively before the index operation and included the following: age (years), body mass index (BMI, kg/m$^2$), preoperative opioid use within 3 months before surgery, smoking history (none, prior, current), diabetes status, drug allergies, number of comorbid conditions, prior ipsilateral hip procedure not including a THA, preoperative PROMs [PRHS, modified Harris Hip Score (MHHS)], and range of motion (ROM) for hip flexion.

### Missing Data

The rates of missing data were MCID (n = 28), preoperative opioids (n = 2), smoking history (n = 2), diabetes (n = 2), comorbidities (n = 2), preoperative health state (n = 26), preoperative mHHS (n = 17), and preoperative ROM (n = 197). ROM was the only variable with greater than 30% missing data and was excluded. Multiple imputation was used for all remaining variables with less than 30% missing data.

### Data Analysis

The detailed methodology used for data analysis has been described extensively in prior studies [13,14]. Briefly, the institutional population of patients was divided in an 80:20 stratified split into training and testing cohorts. Recursive feature selection with random forest algorithms was used to determine the subset of variables used for final algorithm development. The following algorithms were developed on the training set: (1) stochastic gradient boosting, (2) random forest, (3) support vector machine, (4) neural network, and (5) elastic net penalized logistic regression. Metrics used for model performance assessment were discrimination (area under the receiver operating curve [AUC]), calibration (calibration plot, intercept, slope), Brier score, and decision curve analysis. The null model Brier score was calculated to benchmark the algorithm performance. Algorithm performance was assessed on cross-validation of the training set and in the independent testing set. Relative variable importance plots were developed to determine the most important variables for the algorithm with the best overall performance. An open access digital web application was developed with the capacity to

**Table 1**
Baseline Characteristics of Study Population, n = 616.

| Variable | n (%) \| Median (IQR) |
|---|---|
| Age | 62 (54–71) |
| BMI | 29.7 (25.5–36.3) |
| Preoperative opioid use | 169 (27.5) |
| Smoking history | |
| None | 440 (71.7) |
| Prior | 103 (16.8) |
| Current | 71 (11.6) |
| Diabetes | 87 (14.2) |
| Drug allergies | 125 (23.6) |
| Three or more comorbidities | 297 (48.4) |
| Prior ipsilateral hip procedure | 112 (18.2) |
| Preoperative health state | 45 (40–70) |
| Preoperative MHHS | 40 (27–52) |
| Preoperative ROM | 90 (90–100) |
| Patient-reported health state MCID | 407 (69.2) |

Frequency of missing data: MCID (n = 28), preoperative opioids (n = 2), smoking history (n = 2), diabetes (n = 2), comorbidities (n = 2), preoperative health state (n = 26), preoperative MHHS (n = 17), and preoperative ROM (n = 197).
BMI, body mass index; MCID, minimal clinically important difference; MHHS, modified Harris Hip Score; ROM, range of motion.

**Table 2**
Algorithm Performance on Cross-Validation of Training Set, n = 616.

| Metric | Stochastic Gradient Boosting | Random Forest | Support Vector Machine | Neural Network | Elastic Net Penalized Logistic Regression |
|---|---|---|---|---|---|
| AUC | 0.92 (0.89, 0.94) | 0.97 (0.96, 0.98) | 0.90 (0.87, 0.92) | 0.92 (0.90, 0.95) | 0.87 (0.84, 0.90) |
| Calibration intercept | −0.15 (−0.38, 0.07) | 0.33 (−1.25, 1.91) | 0.52 (−1.03, 2.06) | −0.29 (−0.75, 0.18) | −0.08 (−0.27, 0.11) |
| Calibration slope | 1.36 (1.08, 1.64) | 4.53 (1.07, 8.00) | 2.46 (−0.06, 4.98) | 1.35 (0.65, 2.05) | 1.32 (0.89, 1.74) |
| Brier score | 0.089 (0.079, 0.099) | 0.054 (0.047, 0.062) | 0.111 (0.097, 0.124) | 0.083 (0.069, 0.097) | 0.122 (0.109, 0.135) |

Data are expressed as mean (95% confidence interval). Null model Brier score = 0.209.
AUC, area under the receiver operating curve.

provide both predictions and explanations at the individual patient level. The Anaconda Distribution (Anaconda, Inc, Austin, TX), R (The R Foundation, Vienna, Austria), RStudio (RStudio, Boston, MA), and Python (Python Software Foundation, Wilmington, DE) were used for data analysis.

## Results

Overall, 616 patients met the inclusion criteria and had a median age of 62 (interquartile range [IQR] 54–71) years. A total of 353 (57.3%) patients were female. Preoperatively, 27.5% (n = 169) had opioid use and 11.6% (n = 71) were current smokers (Table 1). The median preoperative PRHS score was 45 (IQR 40–70) and the median preoperative mHHS was 40 (IQR 27–52). In all, 69.2% (n = 407) achieved MCID on PRHS scores at 2 years postoperatively.

Variables identified for algorithm development through recursive feature selection were preoperative PRHS, BMI, age, drug allergies, preoperative opioid use, smoking history, prior ipsilateral hip surgery not including a THA, and diabetes.

Cross-validation of the training set (n = 433) resulted in AUC ranging from 0.87 to 0.92, calibration intercept ranging from −0.29 to 0.52, calibration slope ranging from 1.32 to 4.53, and Brier score ranging from 0.054 to 0.122 (Table 2). In comparison, the null model Brier score was 0.209.

Algorithm performance assessment in the testing set (n = 183) resulted in AUC ranging from 0.87 to 0.97, calibration intercept ranging from −0.12 to −0.05, calibration slope ranging from 0.47 to 1.45, and Brier score ranging from 0.054 to 0.122. In comparison, the null model Brier score was 0.208 (Table 3). The null model Brier score was 0.208. The algorithm with the best performance was the random forest with AUC 0.97, calibration intercept −0.05, calibration slope 1.45, and Brier score 0.054. The most important factors for prediction of MCID were preoperative health state, BMI, age, and opioid use (Fig. 1). Decision curve analysis of the random forest algorithm showed that the algorithm resulted in greater net benefit than the default strategies of changes for all patients, for no patients, or changes based on preoperative PRHS alone (Fig. 2). The final algorithm was deployed as an open access digital application and an example of the individual patient-specific explanations included in the application are shown in Figure 3.

## Discussion

The principle findings of the current study are as follows: (1) the random forest algorithm performed well on discrimination, calibration, Brier score, and decision curve analysis when evaluated on the independent testing set for predicting patients who will achieve clinically meaningful improvement for the PRHS; (2) the most important features for prediction were preoperative PRHS score, BMI, age, and preoperative opioid use; and (3) 69.2% of patients achieved clinically meaningful improvement for the PRHS after THA.

The current study assessed the performance of 5 machine learning algorithms, with the random forest approach conferring the greatest predictive value and performance. The factors selected in the current study by recursive feature selection with random forest algorithms to predict patients with the highest probability of achieving clinically significant outcome improvement for the PRHS are in accordance with prior studies that have investigated the predictive potential of preoperative risk factors. Indeed, both age and BMI have been shown to influence outcome expectations after THA [15]. Bonner et al [16] reported that preoperative opioid use was significantly associated with lower mental health scores in patients undergoing primary THA. Wood et al [17] performed a prospective cohort study of 463 total joint arthroplasty patients and found that the most important predictor of pain catastrophizing, anxiety, and depression after total joint arthroplasty was poor subjective health and function. Furthermore, they found that at-risk patients included those who generally reported good function, which is in accordance with the finding in the current study that a greater preoperative PRHS is a strong predictor of failing to achieve clinically significant outcome improvement on the PRHS. Therefore, it is plausible that the factors identified in this study are intricately related to the PRHS and influence the propensity for a clinically significant outcome. These factors should be considered preoperatively—especially those which are modifiable such as BMI and opioid use—in order to optimize a patient's perception of their health state and decrease the risk of poor outcome and readmission. Future studies should seek to further investigate the

**Table 3**
Algorithm Performance in Independent Testing Set (95% Confidence Interval), n = 183.

| Metric | Stochastic Gradient Boosting | Random Forest | Support Vector Machine | Neural Network | Elastic Net Penalized Logistic Regression |
|---|---|---|---|---|---|
| AUC | 0.94 (0.90, 0.97) | 0.97 (0.94, 0.99) | 0.87 (0.80, 0.92) | 0.92 (0.80, 0.95) | 0.87 (0.79, 0.92) |
| Calibration intercept | −0.12 (−0.56, 0.33) | −0.05 (−0.56, 0.46) | −0.06 (−0.45, 0.34) | −0.08 (−0.70, 0.54) | −0.10 (−0.50, 0.31) |
| Calibration slope | 1.38 (0.96, 1.81) | 1.45 (1.00, 1.89) | 1.13 (0.79, 1.46) | 0.47 (0.29, 0.60) | 1.00 (0.68, 1.31) |
| Brier score | 0.083 (0.061, 0.112) | 0.054 (0.036, 0.082) | 0.127 (0.101, 0.164) | 0.082 (0.056, 0.122) | 0.122 (0.097, 0.160) |

Null model Brier score = 0.208.
AUC, area under the receiver operating curve.

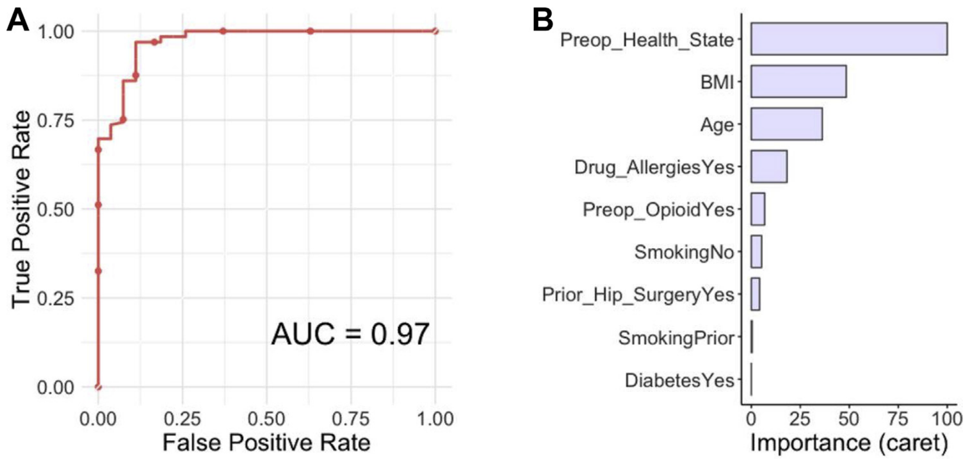


**Fig. 1.** (A) Discrimination of random forest algorithm and (B) relative variable importance testing set (n = 183). AUC, area under the receiver operating curve; BMI, body mass index.

relationships between the PRHS and functional and adverse outcomes.

The majority of patients in this study achieved clinically significant outcome for the PRHS; however, 30.8% of these patients failed to do so. This proportion of MCID achievement is comparable to previously published studies, suggesting the validity of our results for this outcome measure. Van der Wees et al [18] retrospectively reviewed a large clinical registry and determined that the MCID ranged from 66.3% to 87% for the Harris Hip Score and 77.2% to 93.1% for the Oxford Hip Score. Likewise, Lyman et al [19] found that the MCID ranged from 67% to 96% for the Hip Disability and Osteoarthritis Outcome Score and Knee Injury and Osteoarthritis Outcome Score measures. Based on the risk factors identified in the current study, it is likely that the majority of these patients were heavier, older, used opioids preoperatively, and perceived their current overall health and state of their hip as relatively normal. The clinical decision-making tool derived from the current study will aid both patients and total joint arthroplasty surgeons in performing shared decision-making preoperatively as to whether a patient is optimized to undergo THA and whether or not they will perceive their postoperative health as good or improved. Inclusion of the visual algorithm explanations with weighted contribution to risk may allow patients to determine if any of the risk factors are modifiable before surgery. Providers might incorporate these algorithms into the screening and counseling appointments for patients and facilitate preoperative conversations, shared decision-making, and multidisciplinary management, and targeted preoperative interventions based on the model predictions may prevent or reduce the proportion of patients who fail to achieve clinically significant outcome. These algorithms may also become embedded into electronic health records and clinical workflow and obviate the need for providers to manually enter in the values required for preoperative prediction.

### *Limitations*

The results of the current study should be interpreted in the context of few limitations. This study is retrospective in design and utilized chart review, which did not allow for a complete dataset to be constructed. Despite this, missing data were limited and all of the prediction models had excellent performance. The predictive models generated in the current study are inherently validated internally, but this analysis does not allow for external validation. Therefore, future studies will need to determine whether the decision-making model in the current study will apply to other patient cohorts. The models constructed also incorporated only a subset of possible features that are collected in total joint arthroplasty registries, and although the variables had significant contributions in overall effect for predicting clinically significant outcome, it remains possible that other variables which have contributions were not included in this study.

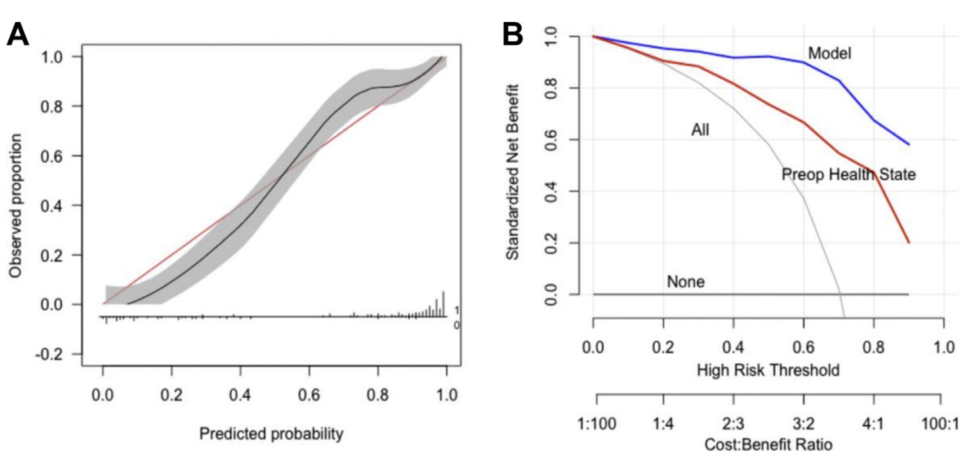


**Fig. 2.** (A) Calibration and (B) decision curve analysis of random forest algorithm testing set (n = 183).

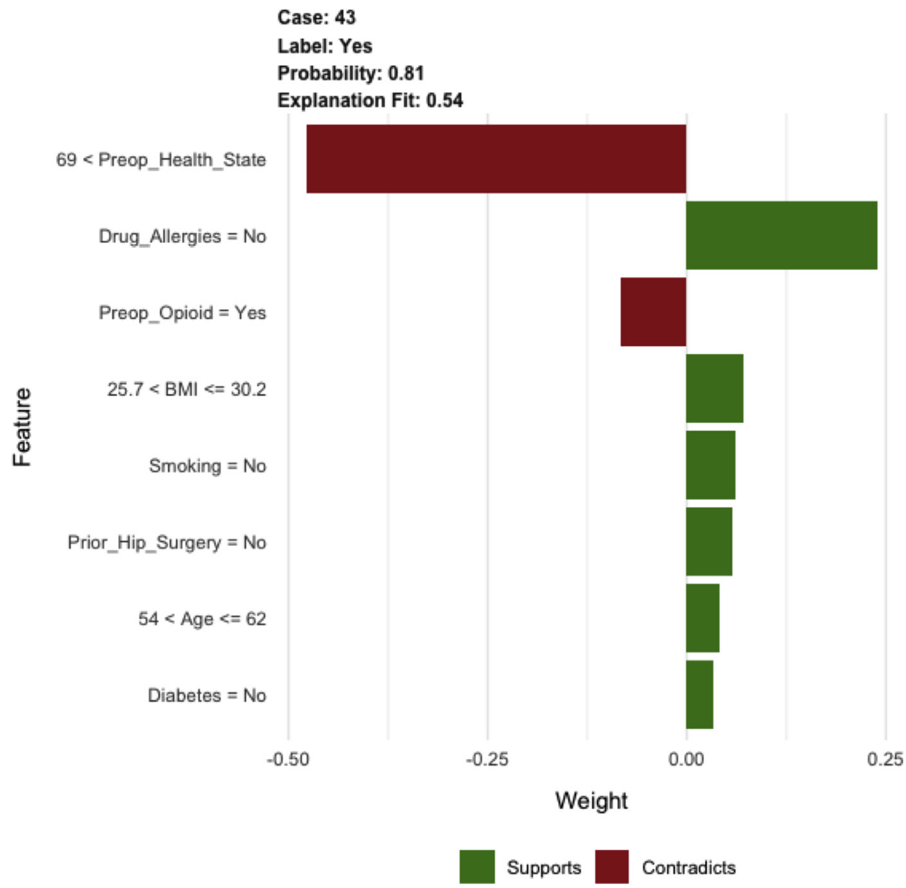


**Fig. 3.** Example of individual patient-level explanation for random forest algorithm predictions.

## Conclusion

The current study generated a clinical decision-making tool based on partially modifiable risk factors for predicting clinically significant outcome improvement after THA. The tool demonstrates excellent discriminative capacity for identifying those at greatest risk for failing to achieve clinically significant improvement in their current health state. External validation is required to further support the utility of this tool in total joint arthroplasty practice.

## References

[1] Kurtz S, Ong K, Lau E, Mowat F, Halpern M. Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030. J Bone Joint Surg Am 2007;89:780–5.
[2] Kunze KN, Akram F, Fuller BC, Zabawa L, Sporer SM, Levine BR. Internal validation of a predictive model for satisfaction after primary total knee arthroplasty. J Arthroplasty 2019;34:663–70.
[3] Kunze KN, Li J, Movassaghi K, Wiggins AB, Sporer SM, Levine BR. Internal validation of a predictive model for complications after total hip arthroplasty. J Arthroplasty 2018;33:3759–67.
[4] Kee J, Mears SC, Edwards PK, Lowry Barnes C. Cost analysis and bundled care of hip and knee replacement. J Surg Orthop Adv 2019;28:241–9.
[5] McLawhorn AS, Buller LT. Bundled payments in total joint replacement: keeping our care affordable and high in quality. Curr Rev Musculoskelet Med 2017;10:370–7.
[6] Palazzo C, Jourdan C, Descamps S, Nizard R, Hamadouche M, Anract P, et al. Determinants of satisfaction 1 year after total hip arthroplasty: the role of expectations fulfilment. BMC Musculoskelet Disord 2014;15:1471–2474.
[7] Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20:1727–36.
[8] Malec JF, Ketchum JM. A standard method for determining the minimal clinically important difference for rehabilitation measures. Arch Phys Med Rehabil 2020;101:1090–4.
[9] Steinhaus ME, Iyer S, Lovecchio F, Khechen B, Stein D, Ross T, et al. Minimal clinically important difference and substantial clinical benefit using PROMIS CAT in cervical spine surgery. Clin Spine Surg 2019;32:392–7.
[10] Nwachukwu BU, Chang B, Beck EC, Neal WH, Movassaghi K, Ranawat AS, et al. How should we define clinically significant outcome improvement on the iHOT-12? HSS J 2019;15:103–8.
[11] Nwachukwu BU, Fields K, Chang B, Nawabi DH, Kelly BT, Ranawat AS. Preoperative outcome scores are predictive of achieving the minimal clinically important difference after arthroscopic treatment of femoroacetabular impingement. Am J Sports Med 2017;45:612–9.
[12] Nwachukwu BU, Chang B, Kahlenberg CA, Fields K, Nawabi DH, Kelly BT, et al. Arthroscopic treatment of femoroacetabular impingement in adolescents provides clinically significant outcome improvement. Arthroscopy 2017;33:1812–8.
[13] Karhade AV, Thio Q, Ogink PT, Bono CM, Ferrone ML, Oh KS, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. Neurosurgery 2019;85:E671–81.
[14] Thio Q, Karhade AV, Ogink PT, Raskin KA, De Amorim Bernstein K, Lozano Calderon SA, et al. Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma? Clin Orthop Relat Res 2018;476:2040–8.
[15] Tilbury C, Haanstra TM, Verdegaal SHM, Nelissen R, de Vet HCW, Vliet Vlieland TPM, et al. Patients' pre-operative general and specific outcome expectations predict postoperative pain and function after total knee and total hip arthroplasties. Scand J Pain 2018;18:457–66.
[16] Bonner BE, Castillo TN, Fitz DW, Zhao JZ, Klemt C, Kwon YM. Preoperative opioid use negatively affects patient-reported outcomes after primary total hip arthroplasty. J Am Acad Orthop Surg 2019;27:e1016–20.
[17] Wood TJ, Thornley P, Petruccelli D, Kabali C, Winemaker M, de Beer J. Preoperative predictors of pain catastrophizing, anxiety, and depression in patients undergoing total joint arthroplasty. J Arthroplasty 2016;31:2750–6.
[18] van der Wees PJ, Wammes JJ, Akkermans RP, Koetsenruijter J, Westert GP, van Kampen A, et al. Patient-reported health outcomes after total hip and knee surgery in a Dutch University Hospital Setting: results of twenty years clinical registry. BMC Musculoskelet Disord 2017;18(1):97–8.
[19] Lyman S, Lee YY, McLawhorn AS, Islam W, MacLean CH. What are the minimal and substantial improvements in the HOOS and KOOS and JR versions after total joint replacement? Clin Orthop Relat Res 2018;476:2432–41.