Follow us @BoneJointRes

Freely available online OPEN ACCESS

BJR

ARTHROPLASTY

Predicting whether patients will achieve minimal clinically important differences following hip or knee arthroplasty

A PERFORMANCE COMPARISON OF MACHINE LEARNING, LOGISTIC REGRESSION, AND PRE-SURGERY PROM SCORES USING DATA FROM NINE GERMAN HOSPITALS

B. Langenberger,
D. Schrednitzki,
A. M. Halder,
R. Busse,
C. M. Pross

From Technische Universität Berlin, Berlin, Germany

Aims

A substantial fraction of patients undergoing knee arthroplasty (KA) or hip arthroplasty (HA) do not achieve an improvement as high as the minimal clinically important difference (MCID), i.e. do not achieve a meaningful improvement. Using three patient-reported outcome measures (PROMs), our aim was: 1) to assess machine learning (ML), the simple pre-surgery PROM score, and logistic-regression (LR)-derived performance in their prediction of whether patients undergoing HA or KA achieve an improvement as high or higher than a calculated MCID; and 2) to test whether ML is able to outperform LR or pre-surgery PROM scores in predictive performance.

Methods

MCIDs were derived using the change difference method in a sample of 1,843 HA and 1,546 KA patients. An artificial neural network, a gradient boosting machine, least absolute shrinkage and selection operator (LASSO) regression, ridge regression, elastic net, random forest, LR, and pre-surgery PROM scores were applied to predict MCID for the following PROMs: EuroQol five-dimension, five-level questionnaire (EQ-5D-5L), EQ visual analogue scale (EQ-VAS), Hip disability and Osteoarthritis Outcome Score-Physical Function Short-form (HOOS-PS), and Knee injury and Osteoarthritis Outcome Score-Physical Function Short-form (KOOS-PS).

Results

Predictive performance of the best models per outcome ranged from 0.71 for HOOS-PS to 0.84 for EQ-VAS (HA sample). ML statistically significantly outperformed LR and pre-surgery PROM scores in two out of six cases.

Conclusion

MCIDs can be predicted with reasonable performance. ML was able to outperform traditional methods, although only in a minority of cases.

Cite this article: *Bone Joint Res* 2023;12(9):512–521.

Keywords: Decision support tool, Machine learning, Method comparison

Article focus

- Applying several machine learning (ML) methods, logistic regression, and pre-surgery PROM scores to predict minimal clinically important differences (MCIDs) in patient-reported outcome measures (PROMs) in a German multicentre dataset of hip and knee arthroplasty patients.

Key messages

- MCIDs can be predicted with fair to good performance.

Correspondence should be sent to Benedikt Langenberger; email: langenberger@tu-berlin.de

doi: 10.1302/2046-3758.129.BJR-2023-0070.R2

Bone Joint Res 2023;12(9):512–521.

- ML outperforms other methods in one-third to half of the cases.
- Pre-surgery PROM scores were the most important predictors.

Strengths and limitations

- Statistically robust comparison of a large variety of methods.
- We used appropriate methods to improve understanding of ML predictions.
- Larger sample size may increase the precision of performance estimates and improve performance.

Introduction

Knee arthroplasty (KA) and hip arthroplasty (HA) are high-volume surgical procedures.[1] A total of 173,625 total knee arthroplasties (TKAs) and 227,851 total hip arthroplasties (THAs) were conducted in Germany in 2020, both ranking among the top 20 procedures with regard to volume in German hospitals.[2] Recently, noticeable increases in KA and HA incidences have been reported in the Organisation for Economic Cooperation and Development (OECD)[3-7] and European countries,[8–10] and rates are projected to further increase dramatically.[4,5,7,11–19]

Nevertheless, high case-volumes do not necessarily indicate high patient-reported satisfaction. It has been reported that up to 30% of patients undergoing HA or KA remain unsatisfied with the outcome.[20–23] Measured by patient-reported outcome measures (PROMs) – that is, standardized questionnaires that measure the patient's health state at a given time – up to 65% of patients do not achieve a minimal clinically important difference (MCID) after HA or KA.[24–27] The MCID is defined as "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management".[28] More easily, it can be defined as "the smallest change that is important to patients",[29] or "smallest benefit of value to patients".[30]

The share of patients failing to achieve a MCID after HA/KA highlights the potential for better decision-making. The success of surgery depends on many individual patient factors, such as the duration and severity of the disease, the extent of perceived pain and discomfort, the use of medication, personal circumstances, concomitant diseases, and expectations.[31–33] As providers' recommendation for surgery can be driven by other factors than clinical guidance alone, e.g. financial incentives,[31] a data-driven decision support tool may be useful. Patients who can be expected to not achieve a MCID may reconsider their choice of treatment, and may be protected from unnecessary risk that comes with surgery.[34] This would improve healthcare systems' resource allocation and also result in fewer disappointed patients.

Machine learning (ML), a sub-branch of artificial intelligence,[35,36] is a promising approach in predicting whether patients achieve MCIDs following HA/KA.[24,26,37–41] In classification tasks, supervised ML can be applied.[36,42,43] ML differs from classical statistical analysis as it can detect non-linearities, interactions, or variable selection itself.[42,44] Logistic regression (LR) was not defined as ML,[45,46] but acted as a comparison method.[47] We further derived predictions using 'simple' pre-surgery PROM scores, an approach that showed promising results in previous research.[40,48]

This study aimed: 1) to assess ML, pre-surgery PROM score, and LR performance in predicting whether patients undergoing HA or KA achieve an improvement as high or higher than a calculated MCID for three PROMs; and 2) to identify if ML is able to outperform LR and/or pre-surgery PROM scores in doing so.

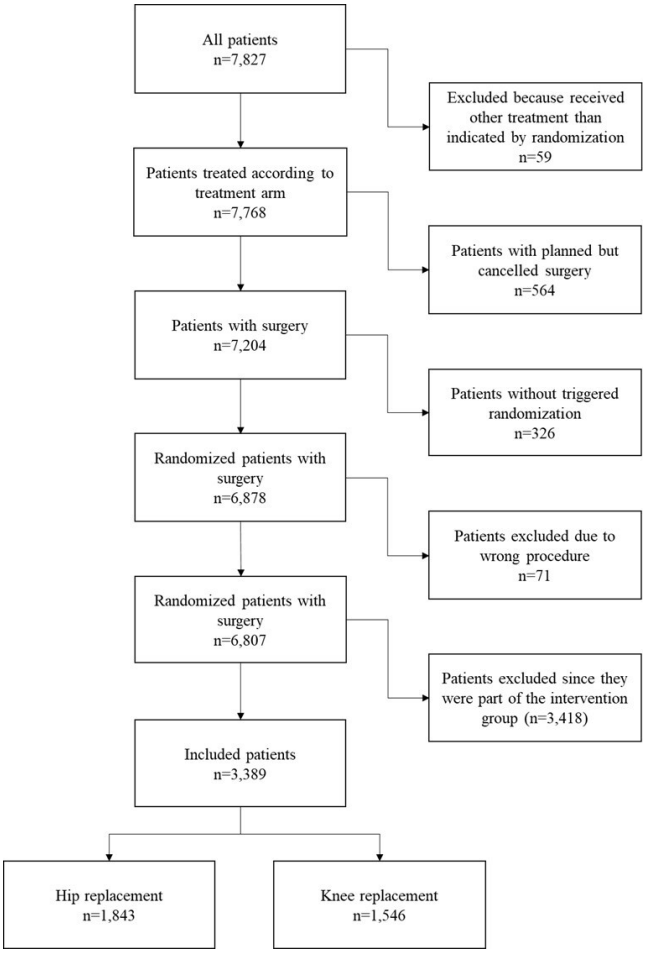

Fig. 1

Flowchart of patient enrolment for this study.

Methods

Data. Data from nine hospitals collected in the German PROMoting Quality study were used.[49] PROMoting Quality was registered under the trial number DRKS00019916 in the German Clinical Trials Register. For this study, only patients from the control group were included since they received treatment as usual. The process of patient selection for this study is illustrated in Figure 1.

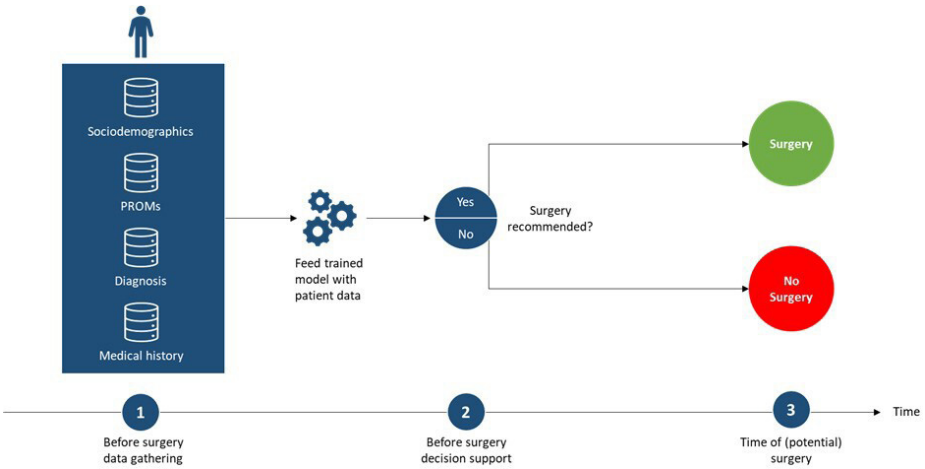

Fig. 2

Graphical illustration of the decision-making support given by the prediction models for practical application. Once relevant data are gathered before surgery (1), trained models are fed with the data and make a prediction (2) about whether surgery is recommended for the respective patient given their input variables. Finally, at the time of (potential) surgery (3), patients recommended to undergo surgery do so, while patients not recommended to be operated do not. PROMs, patient-reported outcome measures.

Out of 7,827 initially recruited patients, 59 were excluded due to receiving another treatment than that indicated by randomization, 564 patients were excluded due to an error in randomization triggering, and 71 received a procedure not indicating HA/KA. After removal of observations from individuals who were part of the intervention group, 1,843 KA and 1,546 HA patients remained in the dataset.

For MCID predictions, we followed Fontana et al[24] and excluded all patients who, mathematically, could not reach a MCID due to a pre-surgery PROM score that was too high, meaning that the addition of the MCID would exceed the scale.

All pre-surgery PROM scores and their dimensions were included as predictors for every outcome. Additionally, age, sex, job status, weight, height, BMI, smoking status, living situation, comorbidities, duration of weekly activity, degree of care dependence, education, and level of physical activity during work/daily routines were included (see Appendix 1 for all variables). After creating dummies for all categorical variables, 198 variables were available for feature selection for HA patients and 203 for KA patients. Differences in variables between HA and KA patients resulted from slightly varying comorbidity profiles and PROM dimensions between both indications.

Missing values and outlier handling. For variables with < 30% missing values, missing values were imputed using missForest,[50] for both categorical and continuous variables. Variables with ≥ 30% missing values were excluded from the analysis. An overview of all variables with missing values is given in Appendix 2.

Patient-reported outcome measures. For MCID calculation, we used only PROMs with evidence about reasonable psychometric properties, namely the generic PROMs EuroQol five-dimension five-level questionnaire (EQ-5D-5L)[51–53] and EQ visual analogue scale (EQ-VAS),[53] as well as the disease-specific Hip disability and Osteoarthritis Outcome Score-Physical Function Short-form (HOOS-PS) and Knee injury and Osteoarthritis Outcome Score-Physical Function Short-form (KOOS-PS).[54–56]

Due to a lack of sufficient validation in arthroplasty patients, Patient-Reported Outcomes Measurement Information System (PROMIS) Fatigue and PROMIS Depression, which were available in the dataset, were not used to determine outcomes, but only as input features.

MCID calculation. We calculated MCIDs using anchor-based methods, as recommended.[57] Patients were asked, "has your health improved as a result of the treatment?" on a Global Rating Scale, which was used as anchor.[58] Possible answers were "worse", "no improvement", "minimal improvement", "improvement", and "great improvement".

The MCID was derived using the change difference (CD) method.[58,59] The CD MCID is calculated as the difference of the mean pre- to post-surgery PROM score change between responders and non-responders. We classified patients who answered "no improvement" on the Global Rating Scale as non-responders, while patients who answered "minimal improvement" were classified as responders.[58]

We used pre-surgery and 12-month post-surgery PROM scores for MCID determination. Previous research found that patient-reported outcomes after HA/KA remain stable from one year after surgery,[60] or even earlier.[61]

When the MCID was smaller in magnitude than the minimal detectable change (MDC), which measures the difference in a given PROM score that is assumed to be a "real" difference rather than only a measurement error,[53] the originally derived MCID was substituted with the MDC.

Prediction methods. ML algorithms that performed well in previous studies,[24,26,37–40] namely an artificial feed-forward

Table I. Mean baseline characteristics (if not otherwise reported) of hip and knee arthroplasty patients (standard deviations in parentheses).

Variable	Hip arthroplasty (n = 1,843)	Knee arthroplasty (n = 1,546)
Age at surgery, yrs	65.99 (10.61)	66.18 (9.4)
BMI, kg/m²	27.87 (5.07)	30.41 (5.68)
HOOS-PS/KOOS-PS baseline	47.1 (16.18)	42.97 (12.05)
HOOS-PS/KOOS-PS outcome	15.19 (14.19)	26.78 (12.87)
EQ-5D-5L baseline	0.6 (0.26)	0.63 (0.25)
EQ-5D-5L outcome	0.87 (0.17)	0.84 (0.19)
EQ-VAS baseline	57.16 (19.72)	58.04 (19.22)
EQ-VAS outcome	73.6 (18.36)	69.93 (18.38)
PROMIS depression baseline	49.84 (8.26)	49.39 (8.15)
PROMIS fatigue baseline	49.23 (9.97)	48.15 (9.54)
Male (fraction)	0.44 (0.5)	0.46 (0.5)
Diabetes (fraction)**	0.09 (0.29)	0.1 (0.3)
Depression (fraction)**	0.06 (0.24)	0.07 (0.25)
Heart disease (fraction)**	0.13 (0.33)	0.12 (0.33)
Back pain (fraction)**	0.21 (0.41)	0.2 (0.4)
At least one hour of weekly activity (fraction)	0.91 (0.28)	0.9 (0.3)
Highest education: high-school or higher (fraction)	0.86 (0.35)	0.82 (0.38)
Working (at least part-time) (fraction)	0.34 (0.47)	0.3 (0.46)
Living in a nursing home (fraction)	0 (0.06)	0.01 (0.08)

*Self-reported (yes/no).
EQ-5D-5L, EuroQol five-dimension five-level questionnaire; HOOS-PS, Hip disability and Osteoarthritis Outcome Score-Physical Function Short Form; KOOS-PS, Knee injury and Osteoarthritis Outcome Score-Physical Function Short Form; PROMIS, Patient-Reported Outcome Measurement Information System; VAS, visual analogue scale.

neural network (NN),[36,42,62] gradient-boosting machine (GBM),[63,64] least absolute shrinkage and selection operator (LASSO) regression,[65–67] ridge regression, elastic net,[65] and random forest (RF)[68] were applied to predict MCIDs. Additionally, LR and pre-surgery PROM scores were applied.[48]

All ML and LR analyses were performed using the h2o package in the statistical software R (R Foundation for Statistical Computing, Austria) and Rstudio (Rstudio, USA). All analyses were run for the KA and HA samples separately. Figure 2 illustrates the data, relevant timepoints, and prediction task of this paper.

Predictive performance measures. Discriminative performance for all applied ML algorithms and LR was assessed using the area under the receiver operating characteristic curve (AUC) as main performance indicator. AUC has a maximum of 1 and a theoretical minimum of 0, while 0.5 indicates predictive performance as good as chance. Performance on AUC is classified as fail (0.5 to 0.59), poor (0.6 to 0.69), fair (0.7 to 0.79), good (0.8 to 0.89), or excellent (0.9 to 1.0).[69] AUC is not attenuated by imbalanced data,[70] and does not rely on a specific sensitivity-specificity trade-off such as other metrics (e.g. Youden Index).[41]

We also report the metric sensitivity, specificity, accuracy, g-mean,[71,72] and Youden Index.[72] Sensitivity, specificity, accuracy, and g-mean were reported at the decision threshold which maximizes the g-mean. For predictions based on pre-surgery PROM scores and the Youden Index itself, sensitivity and specificity were set to maximize the Youden Index.[48]

Further, we report model calibration[73,74] on unforeseen test data,[24] namely the Brier Score[75,76] calibration slope and calibration intercept.[73] Calibration slope and intercept could not be calculated for pre-surgery PROM score predictions, as predicted probabilities were always 0 or 1, and log-odds of predicted probabilities as necessary for calculating calibration slope and intercept could not be derived.[77] Also, 95% asymptotic confidence intervals (CIs) were derived and reported for all performance indicators.[78] AUC comparisons and CIs were derived using the method of Delong et al,[79] with significance set at the level of 5%. It should be noted that although CIs may overlap, AUCs may still turn out to be statistically significantly different based on the test by Delong et al.[79,80] Therefore, when we write that one model outperforms another, we are referring to the fact that the model performs statistically significantly better than another model based on this test.[79]

Data preparation and model selection. The dataset was randomly split into 80% training and 20% test data. Random forest feature selection[81] was applied for each PROM and sample. For all ML algorithms, several hyperparameters were varied in order to select the best possible specification for each algorithm.[42] Hyperparameter tuning was done with fivefold cross-validation (CV)[42] based on the training dataset using grid search.[82] The selected hyperparameters for each model for both KA and HA can be found in Appendix 3. For all ML algorithms, after parameter tuning and performance evaluation, the best-performing specification was selected. All methods were run on the test dataset for final performance assessment and comparison.

Variable importance and explanation. Variable importance was reported using Shapley Additive exPlanations (SHAP) analysis.[83,84] SHAP analysis is a game theory-based approach that ranks variables regarding their influence on different models' predicted probabilities, and facilitates explanations for which values for each variable drive predictions to either increase or decrease.[83,85] Partial dependence plots were used to illustrate the predicted class probability given the pre-surgery PROM scores.[42]

Results

Summary statistics and MCID values. The mean age across both HA and KA patients was approximately 66 years, and a slight majority of individuals were female. Mean BMI was higher in the KA sample (30.41 kg/m²) than in the HA sample (27.87 kg/m²). At 12 months post-surgery, patients in both samples had improved on all scores where

Table II. Results of minimal clinically important difference calculation for the hip arthroplasty and knee arthroplasty samples.

Variable	EQ-5D-5L	EQ-VAS	KOOS-PS
Knee arthroplasty (n = 1,546)*			
MCID	0.20	5.86†	-5.06
MDC	0.10	5.86	-3.67
Share of patients who reached a MCID, %	64.88	64.94	81.76
Share of patients who mathematically could not reach a MCID, %‡	6.99	0.97	0.13
Share of patients who reached a MCID where mathematically possible, %	64.94	64.88	81.76
Hip arthroplasty (n = 1,843)*			
MCID	0.17	7.81	-10.01
MDC	0.10	6.01	-9.42
Share of patients who reached a MCID, %	58.00	66.36	90.56
Share of patients who mathematically could not reach a MCID, %‡	21.38	1.30	0.76
Share of patients who reached a MCID where mathematically possible, %	66.36	58.00	90.56

*Sample size before exclusion of patients who could not reach a MCID.
†MCID values were substituted with MDC values, since in these cases the derived MDC was greater than the derived MCID.
‡In these cases, a MCID could not be reached because the MCID value added (EQ-VAS; EQ-5D-5L) to/subtracted (HOOS-PS/KOOS-PS) from the pre-surgery PROM score extended the PROM's scale. Patients who were not able to reach a MCID were excluded from further analysis.
EQ-5D-5L, EuroQol five-dimension five-level; KOOS-PS, Knee injury and Osteoarthritis Outcome Score-Physical Function Shot Form; MCID, minimal clinically important difference; MDC, minimal detectable change; PROM, patient-reported outcome score; VAS, visual analogue scale.

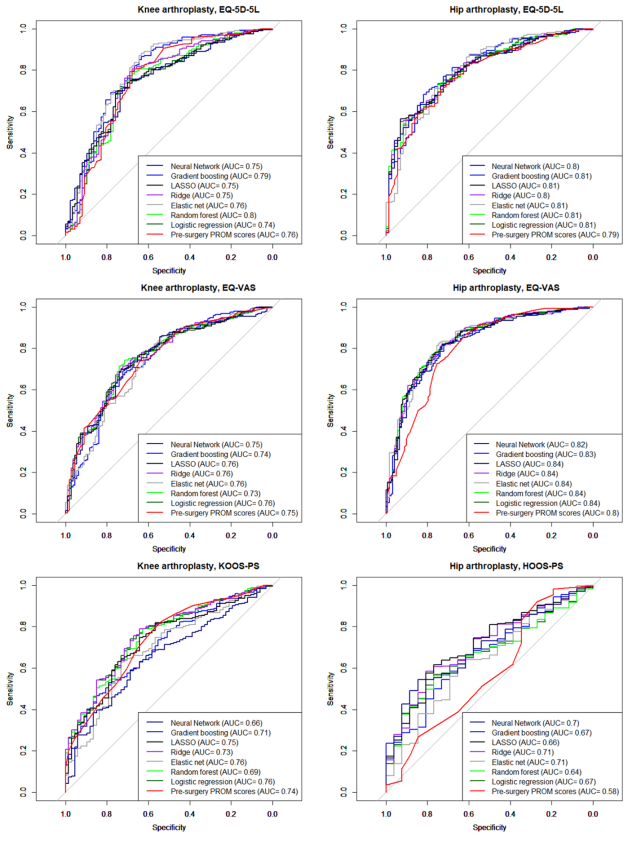

Fig. 3

Receiver operating curves for all models, indications, and patient-reported outcome scores (PROMs). AUC, area under the receiver operating characteristic curve; EQ-5D-5L, EuroQol five-dimension five-level questionnaire; HOOS-PS, Hip disability and Osteoarthritis Outcome Score-Physical Function Short Form; KOOS-PS, Knee injury Osteoarthritis Outcome Score-Physical Function Short Form; LASSO, least absolute shrinkage and selection operator; VAS, visual analogue scale.

MCIDs were calculated. The drop in HOOS-PS scores after surgery was larger than the drop in KOOS-PS scores. Both groups improved substantially on EQ-5D-5L and EQ-VAS, with HA patients achieving slightly larger improvements (Table I).

MCIDs for EQ-5D-5L were 0.20 (KA) and 0.17 (HA), for EQ-VAS 3.27 (KA) and 7.81 (HA), for KOOS-PS -5.06, and for HOOS-PS -10.01 (Table II). The percentages of patients who were mathematically able to reach a MCID varied across PROMs (Table II). While only 0.13% (n = 2) of patients were mathematically unable to reach a MCID in the KA sample for KOOS-PS, 21.38% (n = 394) were unable to reach a MCID in the HA sample for EQ-5D-5L. The share of patients who reached a MCID ranged from 58.00% (n = 840) for EQ-VAS (HA) to 90.56% (n = 1,312) for HOOS-PS (Table II).

Machine learning, logistic regression, and pre-surgery PROM predictive performance. Performance of grid search selected models on training data with fivefold cross-validation was reported in Appendix 4 for all indications and PROMs. Tuning parameters for the selected models are presented in Appendix 3. After training, the selected models were applied to the test dataset for performance assessment (see Figure 2 for receiver operating curves).

The performance[69] of the best models for each outcome ranged between fair (i.e. AUC between 0.7 and 0.8; for knee arthroplasty: EQ-VAS, KOOS-PS; for hip arthroplasty: HOOS-PS) and good (i.e. 0.8 ≤ AUC < 0.9; knee arthroplasty: EQ-5D-5L; hip arthroplasty: EQ-5D-5L, EQ-VAS). In all cases, a ML algorithm was the best-performing model (see Table III and Figure 3).

Table III. Performance assessment of all selected models on unforeseen test data.

Variable	Neural network	Gradient boosting	LASSO	Ridge	Elastic net	Random forest	Logistic regression	Pre-surgery PROM scores
Knee arthroplasty								
EQ-5D-5L (n = 288) AUC (95% CI)	0.76 (0.7 to 0.81)	0.79 (0.74 to 0.84)	0.75 (0.69 to 0.8)	0.75 (0.69 to 0.81)	0.76 (0.7 to 0.81)	0.80 (0.74 to 0.85)*	0.74 (0.68 to 0.8)	0.76 (0.7 to 0.81)
EQ-VAS (n = 307), AUC (95% CI)	0.73 (0.67 to 0.78)	0.74 (0.69 to 0.8)	0.76 (0.71 to 0.82)	0.76 (0.7 to 0.81)	0.76 (0.71 to 0.82)*	0.73 (0.68 to 0.79)	0.76 (0.7 to 0.81)	0.75 (0.7 to 0.81)
KOOS-PS (n = 309), AUC (95% CI)	0.68 (0.62 to 0.75)	0.71 (0.64 to 0.77)	0.75 (0.69 to 0.81)	0.73 (0.67 to 0.79)	0.76 (0.7 to 0.82)*	0.69 (0.63 to 0.76)	0.76 (0.7 to 0.81)	0.74 (0.68 to 0.8)
Hip arthroplasty								
EQ-5D-5L (n = 290), AUC (95% CI)	0.8 (0.75 to 0.86)	0.81 (0.76 to 0.86)*	0.81 (0.76 to 0.86)	0.8 (0.75 to 0.85)	0.81 (0.76 to 0.86)	0.81 (0.75 to 0.86)	0.81 (0.76 to 0.86)	0.79 (0.73 to 0.84)
EQ-VAS (n = 364), AUC (95% CI)	0.82 (0.78 to 0.86)	0.83 (0.79 to 0.87)	0.84 (0.8 to 0.88)*	0.84 (0.8 to 0.88)	0.84 (0.8 to 0.88)	0.84 (0.8 to 0.88)	0.84 (0.8 to 0.88)	0.8 (0.75 to 0.84)
HOOS-PS (n = 366), AUC (95% CI)	0.71 (0.65 to 0.76)	0.67 (0.62 to 0.72)	0.66 (0.61 to 0.72)	0.71 (0.66 to 0.76)*	0.71 (0.65 to 0.76)	0.64 (0.58 to 0.69)	0.67 (0.61 to 0.72)	0.58 (0.47 to 0.68)

*Best-performing model (sometimes identified using further decimal digits than those shown in the table).

Table IV. Statistical difference analysis between different areas under the receiver operating characteristic curve of the best machine learning and non-machine learning method.

PROM	Best ML model	AUC	Comparison 1		Comparison 2	
			Logistic regression (AUC)	p-value*	Pre-surgery PROM scores (AUC)	p-value*
Knee arthroplasty						
EQ-5D-5L	RF	0.80	0.74	0.012‡	0.76	0.052†
EQ-VAS	Elastic net	0.76	0.76	0.401	0.75	0.519
KOOS-PS	Elastic net	0.76	0.76	0.186	0.74	0.355
Hip arthroplasty						
EQ-5D-5L	GBM	0.81	0.81	0.745	0.79	0.242
EQ-VAS	LASSO	0.84	0.84	0.597	0.80	0.034‡
HOOS-PS	Ridge	0.71	0.67	0.017‡	0.58	0.011‡

*p-value for statistical difference of the AUCs of the compared models.
†Indicates statistical significance at the 10% level.
‡Indicates statistical significance at the 5% level.
§Indicates statistical significance at the 1% level.
AUC, area under the curve; EQ-5D-5L, EuroQol five-dimension five-level questionnaire; GBM, gradient-boosting model; HOOS-PS, Hip disability and Osteoarthritis Outcome Score-Physical Function Short Form; KOOS-PS, Knee injury and Osteoarthritis Outcome Score-Physical Function Short Form; LASSO, least absolute shrinkage and selection operator; ML, machine learning; PROM, patient-reported outcome measure; RF, random forest; VAS, visual analogue scale.

Statistical difference testing of AUCs between the best ML model and LR or pre-surgery PROM scores is reported in Table IV.[79]

Statistically significant AUC differences between the best-performing ML model and pre-surgery PROM scores at the 5% level could be identified in two cases, namely for EQ-VAS and HOOS-PS in the HA sample. ML statistically significantly outperformed LR for EQ-5D-5L in the KA sample and for HOOS-PS in the HA sample (Table IV).

SHAP analysis. SHAP analysis for the top ten features was performed for both HA and KA patients based on the GBM (Figure 4).

Red dots in Figure 4 indicate high variable values, and positive x-axis values indicate an increased chance of achieving a MCID. For all PROMs and patient samples, the pre-surgery PROM score of the outcome PROM was ranked as the most important feature. Therefore, better health (high EQ-VAS or EQ-5D-5L/low HOOS-PS/KOOS-PS score) was associated with a lower probability of achieving a MCID.

Further important variables were other PROM scores (and subdimensions) as well as self-reported back pain ("PQ_back") in all cases, BMI and age at surgery in four cases, and height (additional to BMI) in three cases. For all of those variables, a higher variables value (e.g. higher BMI) was associated with decreased likelihood of achieving a MCID.

Partial dependence plots visualize how the probability of achieving a MCID (y-axis) changes when pre-surgery PROM scores change (along the x-axis) for the respective PROM. We observe that, for all PROMs, there seems to be an indeterminate cut-off point after which the probability of achieving a MCID declines steeply (Figure 5).

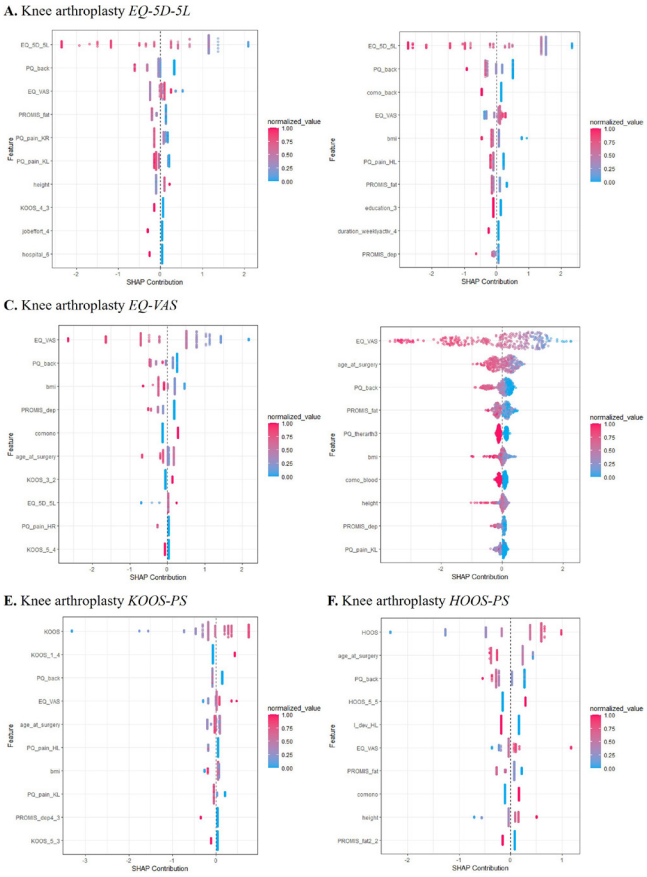

Fig. 4

Shapley Additive exPlanations (SHAP) analysis results for knee arthroplasty (KA) and hip arthroplasty (HA) patients and all patient-reported outcome measures (PROMs). Numbers in PROM names (e.g. KOOS_3_2) represent dummies for response options (e.g. response option 2 in KOOS_3 is KOOS_3_2) and the domain of the PROM (i.e. the third domain in KOOS is KOOS_3_2). EQ-5D-5L, EuroQol five-dimension five-level questionnaire; EQ-VAS, EuroQol visual analogue scale; HOOS-PS, Hip disability and Osteoarthritis Outcome Score-Physical Function Short Form; KOOS-PS, Knee injury and Osteoarthritis Outcome Score-Physical Function Short Form; PQ_back, self-reported back pain; PROMIS, patient-reported outcome measurement information system.

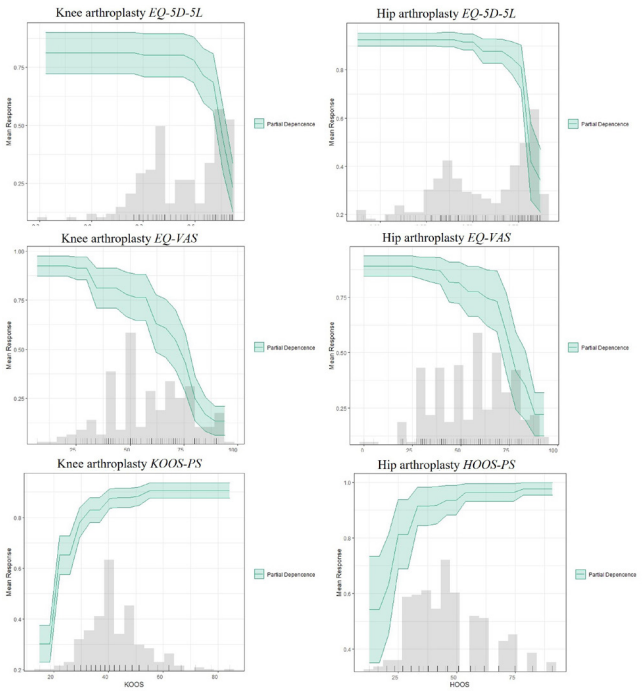

Fig. 5

Partial dependence plots for hip and knee arthroplasty patients and all patient-reported outcome measures. EQ-5D-5L, EuroQol five-dimension five-level questionnaire; EQ-VAS, EuroQol visual analogue scale; HOOS-PS, Hip disability and Osteoarthritis Outcome Score-Physical Function Short Form; KOOS-PS, Knee injury and Osteoarthritis Outcome Score-Physical Function Short Form.

Discussion

This study was the first to make MCID predictions in a German hip and knee arthroplasty sample. It found that ML outperformed both LR and the pre-surgery PROM scores in two out of six cases.

Our findings were partly in line with Zhang et al,[40] who found that pre-surgery PROM scores performed equally as well as ML. In cases where pre-surgery PROM scores perform equally as well as other methods, their application to MCID prediction may likely yield superior clinician and patient adherence to data-driven decision support, due to intuitive interpretation.

The mainly robust performance of LR was in line with some previous evidence.[24,37,38,47] LR did not perform worse than ML in four cases, but there were two cases in which ML outperformed LR. Fontana et al[24] also reported that ML outperformed LR. The present study highlights the relevance of comparing ML models with classical prediction approaches.[40,41] Some previous studies lacked a proper comparison, and may therefore have overemphasized the utility of ML in this research question.[24,26,39]

We further tested whether balancing the data improves the predictive performance of the models,[86] but found that this was not the case. In line with previous evidence,[24,26,37–40] SHAP analysis confirmed that pre-surgery PROM scores were major drivers of MCID prediction for all outcomes and samples. As per previous studies, we found some evidence that lower age[24,39] and lower BMI[24,26] were associated with a better chance of achieving a MCID.

In contrast to Kunze et al,[26] our models did not demonstrate 'excellent' performance for EQ-VAS, even though the sample size was comparable. We argue that the results of studies showing extremely high AUC values should be interpreted with caution if they do not report whether patients who were mathematically unable to reach a MCID were excluded.[26,40] When we included patients who could not reach a MCID, to see how this affected our results, we observed substantially higher AUC values.

Where comparable to previous evidence, our derived MCIDs for EQ-5D-5L, EQ-VAS, and HOOS-PS tended to be lower.[53,87,88] The fraction of patients meeting the KOOS-PS MCID was higher than in another study,[39] and the fraction of patients achieving a MCID on EQ-VAS was remarkably

close to Kunze et al.[26] Although our MCIDs tended to be smaller than in previous studies, we are confident that the MCIDs reflect 'true' differences in changes in PROM scores, and not just measurement error. That is because we compared (and adjusted in one case) the MCIDs to the MDCs (see Methods section). The difference in MCID compared to previous studies may have arisen due to the study sample, the MCID calculation method, and the anchor.

This study comes with some limitations. First, the MCID calculation is unstandardized, and different approaches will yield different results. Second, larger sample sizes are required to derive more precise AUC estimates (see CIs in Table III). Third, the study does not confirm which PROM, or combination of PROMs, is most important for patients undergoing hip or knee arthroplasty. When being used in shared decision-making, it must be defined which (bundle of) PROM(s) is relevant for patients. When a decision support tool predicts that a patient may improve on one PROM and not on another, the consequence remains unclear. This question is of high practical relevance and must be addressed in future research.

In summary, we found that the best models for each outcome performed 'fair' to 'good', according to the definition of Hosmer and Lemeshow, in predicting MCIDs for hip and knee arthroplasty patients,[69] depending on the PROM and subsample under consideration. ML outperformed LR and pre-surgery PROM scores as prediction tool alternatives in two out of six cases, and never performed worse than the other methods. No algorithm consistently performed as the best in all cases. Different ML algorithms should be compared in practice to identify the best for the application at hand. Additional research on the optimal set of PROMs for decision-making is required.

Supplementary material

Tables showing an overview of the complete set of variables as well as the variables selected by the random forest, missing values, tuning parameters, and all discrimination and calibration metrics for training and performance assessment.

References

1. **OECD and European Union**. *Health at a Glance: Europe 2020: OECD*. 2020.
2. **No authors listed**. Statistisches Bundesamt. Gesundheit: Fallpauschalenbezogene Krankenhausstatistik (DRG-Statistik) Operationen Und Prozeduren Der Vollstationären Patientinnen Und Patienten in Krankenhäusern (4-Steller) 2020, 2021. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Krankenhaeuser/Publikationen/Downloads-Krankenhaeuser/operationen-prozeduren-5231401207014.html (date last accessed 24 July 2023).
3. **OECD**. *Health at a Glance 2015: OECD Indicators*. OECD Publishing, 2015.
4. **Pilz V**, **Hanstein T**, **Skripitz R**. Projections of primary hip arthroplasty in Germany until 2040. *Acta Orthop*. 2018;89(3):308–313.
5. **Klug A**, **Gramlich Y**, **Rudert M**, **et al**. The projected volume of primary and revision total knee arthroplasty will place an immense burden on future health care systems over the next 30 years. *Knee Surg Sports Traumatol Arthrosc*. 2021;29(10):3287–3298.
6. **Kurtz SM**, **Ong KL**, **Lau E**, **Bozic KJ**. Impact of the economic downturn on total joint replacement demand in the United States: updated projections to 2021. *J Bone Joint Surg Am*. 2014;96-A(8):624–630.
7. **Inacio MCS**, **Graves SE**, **Pratt NL**, **Roughead EE**, **Nemes S**. Increase in total joint arthroplasty projected from 2014 to 2046 in Australia: A conservative local model with international implications. *Clin Orthop Relat Res*. 2017;475(8):2130–2137.
8. **Kurtz SM**, **Ong KL**, **Lau E**, **et al**. International survey of primary and revision total knee replacement. *Int Orthop*. 2011;35(12):1783–1789.
9. **Leitner L**, **Türk S**, **Heidinger M**, **et al**. Trends and economic impact of hip and knee arthroplasty in Central Europe: Findings from the Austrian National Database. *Sci Rep*. 2018;8(1):4707.
10. **Le Stum M**, **Gicquel T**, **Dardenne G**, **Le Goff-Pronost M**, **Stindel E**, **Clavé A**. Total knee arthroplasty in France: Male-driven rise in procedures in 2009-2019 and projections for 2050. *Orthop Traumatol Surg Res*. 2022;103463.
11. **Culliford D**, **Maskell J**, **Judge A**, **et al**. Future projections of total hip and knee arthroplasty in the UK: results from the UK Clinical Practice Research Datalink. *Osteoarthritis Cartilage*. 2015;23(4):594–600.
12. **Rupp M**, **Lau E**, **Kurtz SM**, **Alt V**. Projections of primary TKA and THA in Germany from 2016 through 2040. *Clin Orthop Relat Res*. 2020;478(7):1622–1633.
13. **Hooper G**, **Lee A-J**, **Rothwell A**, **Frampton C**. Current trends and projections in the utilisation rates of hip and knee replacement in New Zealand from 2001 to 2026. *N Z Med J*. 2014;127(1401):82–93.
14. **Nemes S**, **Gordon M**, **Rogmark C**, **Rolfson O**. Projections of total hip replacement in Sweden from 2013 to 2030. *Acta Orthop*. 2014;85(3):238–243.
15. **Nemes S**, **Rolfson O**, **W-Dahl A**, **et al**. Historical view and future demand for knee arthroplasty in Sweden. *Acta Orthop*. 2015;86(4):426–431.
16. **Patel A**, **Pavlou G**, **Mújica-Mota RE**, **Toms AD**. The epidemiology of revision total knee and hip arthroplasty in England and Wales: A comparative analysis with projections for the United States. A study using the National Joint Registry dataset. *Bone Joint J*. 2015;97-B(8):1076–1081.
17. **Singh JA**, **Yu S**, **Chen L**, **Cleveland JD**. Rates of total joint replacement in the United States: Future projections to 2020-2040 using the national inpatient sample. *J Rheumatol*. 2019;46(9):1134–1140.
18. **Sloan M**, **Premkumar A**, **Sheth NP**. Projected volume of primary total joint arthroplasty in the U.S., 2014 to 2030. *J Bone Joint Surg Am*. 2018;100-A(17):1455–1460.
19. **Kumar A**, **Tsai W-C**, **Tan T-S**, **Kung P-T**, **Chiu L-T**, **Ku M-C**. Temporal trends in primary and revision total knee and hip replacement in Taiwan. *J Chin Med Assoc*. 2015;78(9):538–544.
20. **Gandhi R**, **Davey JR**, **Mahomed NN**. Predicting patient dissatisfaction following joint replacement surgery. *J Rheumatol*. 2008;35(12):2415–2418.
21. **Price AJ**, **Alvand A**, **Troelsen A**, **et al**. Knee replacement. *Lancet*. 2018;392:1672–1682.
22. **Halawi MJ**, **Jongbloed W**, **Baron S**, **Savoy L**, **Williams VJ**, **Cote MP**. Patient dissatisfaction after primary total joint arthroplasty: The patient perspective. *J Arthroplasty*. 2019;34(6):1093–1096.
23. **Bourne RB**, **Chesworth BM**, **Davis AM**, **Mahomed NN**, **Charron KDJ**. Patient satisfaction after total knee arthroplasty: who is satisfied and who is not? *Clin Orthop Relat Res*. 2010;468(1):57–63.
24. **Fontana MA**, **Lyman S**, **Sarker GK**, **Padgett DE**, **MacLean CH**. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clin Orthop Relat Res*. 2019;477(6):1267–1279.
25. **van der Wees PJ**, **Wammes JJG**, **Akkermans RP**, **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 musculoskeletal disorders 2017;18:97. *BMC Musculoskelet Disord*. 2017;18(1):97.
26. **Kunze KN**, **Karhade AV**, **Sadauskas AJ**, **Schwab JH**, **Levine BR**. Development of machine learning algorithms to predict clinically meaningful improvement for the patient-reported health state after total hip arthroplasty. *J Arthroplasty*. 2020;35(8):2119–2123.
27. **Quintana JM**, **Aguirre U**, **Barrio I**, **Orive M**, **Garcia S**, **Escobar A**. Outcomes after total hip replacement based on patients' baseline status: what results can be expected? *Arthritis Care Res (Hoboken)*. 2012;64(4):563–572.
28. **Jaeschke R**, **Singer J**, **Guyatt GH**. Measurement of health status. *Controlled Clinical Trials*. 1989;10(4):407–415.
29. **Riddle DL**, **Stratford PW**, **Binkley JM**. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 2. *Phys Ther*. 1998;78(11):1197–1207.
30. **McGlothlin AE**, **Lewis RJ**. Minimal clinically important difference: defining what really matters to patients. *JAMA*. 2014;312(13):1342–1343.
31. **Papanicolas I**, **McGuire A**. Do financial incentives trump clinical guidance? Hip replacement in England and Scotland. *J Health Econ*. 2015;44:25–36.

32. **Mota REM**, **Tarricone R**, **Ciani O**, **Bridges JFP**, **Drummond M**. Determinants of demand for total hip and knee arthroplasty: a systematic literature review. *BMC Health Serv Res*. 2012;12:225.
33. **Podmore B**, **Hutchings A**, **van der Meulen J**, **Aggarwal A**, **Konan S**. Impact of comorbid conditions on outcomes of hip and knee replacement surgery: a systematic review and meta-analysis. *BMJ Open*. 2018;8(7):e021784.
34. **Mujica-Mota RE**, **Watson LK**, **Tarricone R**, **Jäger M**. Cost-effectiveness of timely *versus* delayed primary total hip replacement in Germany: A social health insurance perspective. *Orthop Rev (Pavia)*. 2017;9(3):7161.
35. **Russell SJ**, **Norvig P**. *Artificial Intelligence: A Modern Approach*. Upper Saddle River, New Jersey, USA: Prentice Hall, 1999.
36. **Russell SJ**, **Norvig P**, **Davis E**, **Edwards D**. *Artificial Intelligence: A Modern Approach*. Pearson, 2016.
37. **Huber M**, **Kurz C**, **Leidl R**. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. *BMC Med Inform Decis Mak*. 2019;19(1):3.
38. **Harris AHS**, **Kuo AC**, **Bowe TR**, **Manfredi L**, **Lalani NF**, **Giori NJ**. Can machine learning methods produce accurate and easy-to-use preoperative prediction models of one-year improvements in pain and functioning after knee arthroplasty? *J Arthroplasty*. 2021;36(1):112–117.
39. **Katakam A**, **Karhade AV**, **Collins A**, **et al**. Development of machine learning algorithms to predict achievement of minimal clinically important difference for the KOOS-PS following total knee arthroplasty. *J Orthop Res*. 2022;40(4):808–815.
40. **Zhang S**, **Lau BPH**, **Ng YH**, **Wang X**, **Chua W**. Machine learning algorithms do not outperform preoperative thresholds in predicting clinically meaningful improvements after total knee arthroplasty. *Knee Surg Sports Traumatol Arthrosc*. 2022;30(8):2624–2630.
41. **Langenberger B**, **Thoma A**, **Vogt V**. Can minimal clinically important differences in patient reported outcome measures be predicted by machine learning in patients with total knee or hip arthroplasty? A systematic review. *BMC Med Inform Decis Mak*. 2022;22(1):18.
42. **Hastie T**, **Tibshirani R**, **Friedman J**. The Elements of Statistical Learning. In: *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009.
43. **Jiang F**, **Jiang Y**, **Zhi H**, **et al**. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230–243.
44. **Boulesteix A-L**, **Schmid M**. Machine learning versus statistical modeling. *Biom J*. 2014;56(4):588–593.
45. **Bracher-Smith M**, **Crawford K**, **Escott-Price V**. Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Mol Psychiatry*. 2021;26(1):70–79.
46. **Garcia EA**, **Haibo He**. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*. ;21(9):1263–1284. 2009
47. **Christodoulou E**, **Ma J**, **Collins GS**, **Steyerberg EW**, **Verbakel JY**, **Van Calster B**. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22.
48. **Berliner JL**, **Brodke DJ**, **Chan V**, **SooHoo NF**, **Bozic KJ**. Can preoperative patient-reported outcome measures be used to predict meaningful improvement in function after TKA? *Clin Orthop Relat Res*. 2017;475(1):149–157.
49. **Kuklinski D**, **Oschmann L**, **Pross C**, **Busse R**, **Geissler A**. The use of digitally collected patient-reported outcome measures for newly operated patients with total knee and hip replacements to improve post-treatment recovery: study protocol for a randomized controlled trial. *Trials*. 2020;21(1):322.
50. **Stekhoven DJ**, **Bühlmann P**. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112–118.
51. **Jin X**, **Al Sayah F**, **Ohinmaa A**, **Marshall DA**, **Johnson JA**. Responsiveness of the EQ-5D-3L and EQ-5D-5L in patients following total hip or knee replacement. Quality of life research an international journal of quality of life aspects of treatment, care and rehabilitation 2019;28:2409–17. *Qual Life Res*. 2019;28(9):2409–2417.
52. **Conner-Spady BL**, **Marshall DA**, **Bohm E**, **Dunbar MJ**, **Noseworthy TW**. Comparing the validity and responsiveness of the EQ-5D-5L to the Oxford hip and knee scores and SF-12 in osteoarthritis patients 1 year following total joint replacement. Quality of life research an international journal of quality of life aspects of treatment, care and rehabilitation 2018;27:1311–22. *Qual Life Res*. 2018;27(5):1311–1322.
53. **Bilbao A**, **García-Pérez L**, **Arenaza JC**, **et al**. Psychometric properties of the EQ-5D-5L in patients with hip or knee osteoarthritis: reliability, validity and responsiveness. Quality of life research an international journal of quality of life aspects of treatment, care and rehabilitation 2018;27:2897–908. *Qual Life Res*. 2018;27(11):2897–2908.
54. **Alviar MJ**, **Olver J**, **Brand C**, **et al**. Do patient-reported outcome measures in hip and knee arthroplasty rehabilitation have robust measurement attributes? A systematic review. *J Rehabil Med*. 2011;43(7):572–583.
55. **Davis AM**, **Perruccio AV**, **Canizares M**, **et al**. Comparative, validity and responsiveness of the HOOS-PS and KOOS-PS to the WOMAC physical function subscale in total joint replacement for osteoarthritis. *Osteoarthritis and Cartilage*. 2009;17(7):843–847.
56. **Harris K**, **Dawson J**, **Gibbons E**. Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty. *Patient Relat Outcome Meas*. 2016;7:101–108.
57. **Mouelhi Y**, **Jouve E**, **Castelli C**, **Gentile S**. How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods. *Health Qual Life Outcomes*. 2020;18(1):136.
58. **Copay AG**, **Subach BR**, **Glassman SD**, **Polly DW**, **Schuler TC**. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J*. 2007;7(5):541–546.
59. **Copay AG**, **Glassman SD**, **Subach BR**, **Berven S**, **Schuler TC**, **Carreon LY**. Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales. *Spine J*. 2008;8(6):968–974.
60. **Galea VP**, **Rojanasopondist P**, **Connelly JW**, **et al**. Changes in patient satisfaction following total joint arthroplasty. *J Arthroplasty*. 2020;35(1):32–38.
61. **Canfield M**, **Savoy L**, **Cote MP**, **Halawi MJ**. Patient-reported outcome measures in total joint arthroplasty: defining the optimal collection window. *Arthroplast Today*. 2020;6(1):62–67.
62. **Schmidhuber J**. Deep learning in neural networks: an overview. *Neural Netw*. 2015;61:85–117.
63. **Ayyadevara VK.** Gradient Boosting Machine. In: Ayyadevara VK, editor. *Pro Machine Learning Algorithms*. Berkeley, California, USA: Apress; 2018, p. 117–134.
64. **Friedman JH**. Greedy function approximation: A gradient boosting machine. *Ann Statist*. 2001;29(5).
65. **Zou H**, **Hastie T**. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*. 2005;67(2):301–320.
66. **Çiftsüren MN**, **Akkol S**. Prediction of internal egg quality characteristics and variable selection using regularization methods: ridge, LASSO and elastic net. *Arch Anim Breed*. 2018;61(3):279–284.
67. **Ogutu JO**, **Schulz-Streeck T**, **Piepho H-P**. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proc*. 2012;6 Suppl 2(Suppl 2):S10.
68. **Breiman L**. Random forest. *Mach Learn*. 2001;45(1):5–32.
69. **Hosmer DW**, **Lemeshow S.** *Applied Logistic Regression*. 2nd ed. New York, New York, USA: John Wiley; 2010.
70. **Jeni LA**, **Cohn JF**, **De La Torre F**. Facing imbalanced data recommendations for the use of performance metrics. *Int Conf Affect Comput Intell Interact Workshops*. 2013;2013:245–251.
71. **Izad Shenas SA**, **Raahemi B**, **Hossein Tekieh M**, **Kuziemsky C**. Identifying high-cost patients using data mining techniques and a small set of non-trivial attributes. *Comput Biol Med*. 2014;53:9–18.
72. **Bekkar M**, **Djemaa HK**, **Alitouche TA**. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*. 2013;10:27–39.
73. **Van Calster B**, **McLernon DJ**, **van Smeden M**, **Wynants L**, **Steyerberg EW**, **Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative**. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.
74. **Steyerberg EW**, **Vickers AJ**, **Cook NR**, **et al**. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138.
75. **Fenlon C**, **O'Grady L**, **Doherty ML**, **Dunnion J**. A discussion of calibration techniques for evaluating binary and categorical predictive models. *Prev Vet Med*. 2018;149:107–114.
76. **Brier GW**. Verification of forecasts expressed in terms of probability. *Mon Wea Rev*. 1950;78(1):1–3.
77. **Huang Y**, **Li W**, **Macheret F**, **Gabriel RA**, **Ohno-Machado L**. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc*. 2020;27(4):621–633.
78. **Wallace IF**, **Berkman ND**, **Watson LR**, **et al**. Screening for speech and language delay in children 5 years old and younger: A systematic review. *Pediatrics*. 2015;136(2):e448–62.
79. **DeLong ER**, **DeLong DM**, **Clarke-Pearson DL**. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*. 1988;44(3):837–845.

80. **Robin X**, **Turck N**, **Hainard A**, **et al**. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77.
81. **Calle ML**, **Urrea V**, **Boulesteix A-L**, **Malats N**. AUC-RF: a new strategy for genomic profiling with random forest. *Hum Hered*. 2011;72(2):121–132.
82. **Liashchynskyi P**, **Liashchynskyi P**. Grid search, random search, genetic algorithm: A big comparison for NAS: arXiv. Cornell University. 2019. https://arxiv.org/abs/1912.06059 (date last accessed 26 July 2023).
83. **Mangalathu S**, **Hwang S-H**, **Jeon J-S**. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Engineering Structures*. 2020;219:110927.
84. **Lundberg S**, **Lee S-I**. A unified approach to interpreting model predictions. Cornell University. 2017. https://arxiv.org/abs/1705.07874 (date last accessed 26 July 2023).
85. **Snider B**, **McBean EA**, **Yawney J**, **Gadsden SA**, **Patel B**. Corrigendum: Identification of variable importance for predictions of mortality from COVID-19 using AI models for Ontario, Canada. *Front Public Health*. 2021;9:759014.
86. **Kaur H**, **Pannu HS**, **Malhi AK**. A systematic review on imbalanced data challenges in machine learning. *ACM Comput Surv*. 2020;52(4):1–36.
87. **Impellizzeri FM**, **Mannion AF**, **Naal FD**, **Hersche O**, **Leunig M**. The early outcome of surgical treatment for femoroacetabular impingement: success depends on how you measure it. *Osteoarthritis Cartilage*. 2012;20(7):638–645.
88. **Paulsen A**, **Roos EM**, **Pedersen AB**, **Overgaard S**. Minimal clinically important improvement (MCII) and patient-acceptable symptom state (PASS) in total hip arthroplasty (THA) patients 1 year postoperatively. *Acta Orthop*. 2014;85(1):39–48.

Author information:
- B. Langenberger, MSc, Research Associate
- R. Busse, MD, MPH, FFPH, Professor
- C. M. Pross, PhD, Senior Research Fellow
  Health Care Management, Technische Universität Berlin, Berlin, Germany.
- D. Schrednitzki, MD, Orthopaedic Surgeon, Senior Physician
- A. M. Halder, MD, Orthopaedic Surgeon, Chief Physician
  Orthopedics, Sana Kliniken Sommerfeld, Kremmen, Germany.

Author contributions:
- B. Langenberger: Data curation, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing.
- D. Schrednitzki: Methodology, Validation, Writing – review & editing.
- A. M. Halder: Validation, Writing – review & editing.
- R. Busse: Project administration, Resources, Validation, Writing – review & editing.
- C. M. Pross: Project administration, Resources, Validation, 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: the study was funded by the Innovation Fund of the German Federal Joint Committee (G-BA) in the stream "Care models with comprehensive and measurable results and process responsibility" under the funding code 01NVF18016.

ICMJE COI statement:
- D. Schrednitzki reports payments for lectures and courses on knee arthroplasty and robotics from Zimmer Biomet, unrelated to this study. R. Busse reports institutional grants from Roche and Stryker, and speaker payments from AbbVie, all of which are unrelated to this study. R. Busse is also involved with the Government Commission on Hospital Reform. A. Halder reports royalties or licenses, speaker payments, and support for attending meetings and/or travel from Zimmer Biomet and DePuy, unrelated to this study. A. Halder is also President of the German Orthopaedic Society (DGOOC) 2022 Board Member European Knee Society. C. Pross is employed by Stryker, and reports stock in Stryker, unrelated to this study.

Data sharing:
- The datasets generated and analyzed in the current study are not publicly available due to data protection regulations. Access to data is limited to the researchers who have obtained permission for data processing. Further inquiries can be made to the corresponding author.

Acknowledgements:
- We thank the PROMoting Quality team at TU Berlin for project planning, management, and general support.

Open access funding:
- The authors acknowledge support by the German Research Foundation and the Open Access Publication Fund of TU Berlin.