Development of Machine-Learning Algorithms to Predict Attainment of Minimal Clinically Important Difference After Hip Arthroscopy for Femoroacetabular Impingement Yield Fair Performance and Limited Clinical Utility

M. H. Pettit, M.B., B.Chir., B.A., S. H. M. Hickman, M.Sci., M.Res., B.A., A. Malviya, Ph.D., F.R.C.S. (Orth.), M.Sc., and V. Khanduja, Ph.D., F.R.C.S. (T&O.), M.A. (Cantab.), M.Sc., M.B.B.S.

Purpose: To determine whether machine learning (ML) techniques developed using registry data could predict which patients will achieve minimum clinically important difference (MCID) on the International Hip Outcome Tool 12 (iHOT-12) patient-reported outcome measures (PROMs) after arthroscopic management of femoroacetabular impingement syndrome (FAIS). And secondly to determine which preoperative factors contribute to the predictive power of these models. **Methods:** A retrospective cohort of patients was selected from the UK's Non-Arthroplasty Hip Registry. Inclusion criteria were a diagnosis of FAIS, management via an arthroscopic procedure, and a minimum follow-up of 6 months after index surgery from August 2012 to June 2021. Exclusion criteria were for non-arthroscopic procedures and patients without FAIS. ML models were developed to predict MCID attainment. Model performance was assessed using the area under the receiver operating characteristic curve (AUROC). **Results:** In total, 1,917 patients were included. The random forest, logistic regression, neural network, support vector machine, and gradient boosting models had AUROC 0.75 (0.68-0.81), 0.69 (0.63-0.76), 0.69 (0.63-0.76), 0.70 (0.64-0.77), and 0.70 (0.64-0.77), respectively. Demographic factors and disease features did not confer a high predictive performance. Baseline PROM scores alone provided comparable predictive performance to the whole dataset models. Both EuroQoL 5-Dimension 5-Level and iHOT-12 baseline scores and iHOT-12 baseline scores alone provided AUROC of 0.74 (0.68-0.80) and 0.72 (0.65-0.78), respectively, with random forest models. **Conclusions:** ML models were able to predict with fair accuracy attainment of MCID on the iHOT-12 at 6-month postoperative assessment. The most successful models used all patient variables, all baseline PROMs, and baseline iHOT-12 responses. These models are not sufficiently accurate to warrant routine use in the clinic currently. **Level of Evidence:** Level III, retrospective cohort design; prognostic study.

Femoroacetabular impingement syndrome (FAIS) is a common cause of hip pain in the young adult, frequently seen within sporting populations, and is a proposed precursor to hip osteoarthritis.[1-3] FAIS diagnoses have risen dramatically over the past 2 decades, as has the frequency of arthroscopic hip surgery for FAIS management.[4,5] Arthroscopic management demonstrates good outcomes with

From St. George's University Hospital, London, United Kingdom (M.H.P.); The Alan Turing Institute, London, United Kingdom (S.H.M.H.); Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, United Kingdom (S.H.M.H.); Newcastle University, Newcastle upon Tyne, United Kingdom (A.M.); and Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom (V.K.).

Received April 24, 2023; accepted September 13, 2023.

Address correspondence to V. Khanduja, Ph.D., F.R.C.S. (T&O.), M.A. (Cantab.), M.Sc., M.B.B.S, Young Adult Hip Service, Addenbrooke's - Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. E-mail: vk279@cam.ac.uk

0749-8063/23540

https://doi.org/10.1016/j.arthro.2023.09.023

significant improvements in patient-reported outcome measures (PROMs), restoration of nascent hip anatomy, and normalization of hip biomechanics.[6-10]

PROMs have increasingly gained recognition for the assessment of surgical outcome.[11] The Warwick consensus agreement on FAIS specified the use of PROMs, including the International Hip Outcome Tool 12 (iHOT-12), to assess treatment outcome. The minimally clinically important difference (MCID) is frequently used to define success. MCID is the smallest perceptible change patients recognize as beneficial, and only 63.4% to 87.5% of patients undergoing arthroscopic management of FAIS achieve MCID for iHOT-12.[11-14] Improving MCID attainment rate following surgery is a key goal for clinicians, and although good outcomes are inherently linked to the technical skill of the operator, this could also be achieved through improved patient selection and other interventions like prehabilitation.[15] Indeed, it is now recognized that machine-learning (ML) techniques can be used to refine and improve care provision in the field through predicting postoperative PROMs.[16]

Previous studies have used ML techniques to predict MCID attainment after hip arthroscopy using baseline patient characteristics as predictive features.[17-19] In these studies, models have performed well, with the cited articles developing models with an area under the receiver operating characteristic curve (AUROC), or accuracy, of >0.80. The best-performing model boasted an AUROC of 0.89. These models are highly accurate at predicting outcome in their sample population; however, across the orthopaedic field, there is as-yet limited evidence that the application of such models yields tangible improvements in patient outcome.[20,21]

The authors believe that barriers exist for the implementation of such models and the subsequent demonstration of their clinical utility. In a recent systematic review, it was found that only 1 of 18 studies predicting outcome after orthopaedic surgery had included external validation.[22] The aforementioned studies using ML to predict outcome in hip arthroscopy have all been performed using a single-surgeon cohort, and one of these models has been externally validated in another center. There is currently no evidence for the development of accurate ML models in a wider dataset, such as a national registry. A further barrier to clinical utility may be that models require a large number of predictive feature variables, hampering real-world clinical utility in the preoperative clinic and decision making.

The aim of this study, therefore, was to determine whether ML techniques developed using registry data could predict which patients would achieve MCID on the iHOT-12 PROM after arthroscopic management of FAIS and secondly to determine which preoperative factors contribute to the predictive power of these models. The hypothesis of this study was that ML techniques utilised on the UK Non-Arthroplasty Hip Registry (NAHR) data would accurately predict which patients were able to achieve MCID on the iHOT-12 and that this could be achieved using the iHOT-12 alone.

Methods

This work was prepared in accordance with Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidance.[23]

Data Source

The NAHR steering committee granted approval for this retrospective cohort study of prospectively collected data. Retrospective data were collected from the NAHR for the period August 1, 2012, to June 1, 2021. Inclusion criteria were patients who had undergone arthroscopy for FAIS pathology and who had 6-month follow-up iHOT-12 overall scores recorded. All femoral, acetabular, and labral procedures carried out through arthroscopic techniques were included. Exclusion criteria were for open procedures and arthroscopic procedures conducted without a recorded FAIS diagnosis. All patients who met the inclusion criteria were selected from the database (Appendix Table 1, available at www.arthroscopyjournal.org).

Outcome Measure

The outcome measure of interest in this study was achievement of MCID of iHOT-12 at 6 months' follow-up. The MCID value used was 13.0, determined in 2 previous studies validating the iHOT-12 questionnaire in English-speaking cohorts.[24,25] The area AUROC was chosen as the primary outcome measure for model predictive performance. AUROC varies from 0.5 (random predictor) to 1 (perfect predictor). Values of greater than 0.9 are considered to represent excellent predictive performance, 0.8-0.9 good, 0.7-0.8 fair, and <0.7 is poor.[26] Secondary outcome measures are included in Appendix Tables 2 and 3 (available at www.arthroscopyjournal.org) and included precision, defined by: $\frac{True\ Positives}{True\ Positives+False\ Positives}$ recall, defined by: $\frac{True\ Positives}{True\ Positives+False\ Negatives}$ and F1 score, defined by: $\frac{2(Precision)(Recall)}{Precision+Recall}$.

Predictive Feature Variables

Preoperative, baseline data for patient characteristics were used as predictive features for outcome prediction. Arthroscopy-specific parameters, for instance, acetabular procedures performed and femoral procedures performed, were not used as predictive variables. These features do not enable a ML-based tool to predict outcome entirely preoperatively, as preoperative plans may vary significantly from procedure performed. The baseline data were grouped across 3 categories: patient demographics (sex, age at index procedure, body mass index [BMI]), disease status (laterality of operation, acetabular cartilage damage, Outerbridge classification,

acetabular cartilage damage zone, primary or revision surgery), and baseline PROM scores (EuroQoL 5-Dimension 5-Level [EQ-5D-5L] consisting of both the EQ-5D-5L descriptive system: overall score and individual questions, and the EQ visual analog scale for impression of overall health, iHOT-12 overall score, and individual questions). Categorical data variables such as sex were converted to numerical values with one-hot encoding, whereas variables that were answered on a Likert scale with text answers (eg, EQ-5D-5L, "I have no problems in walking about") were converted to a numerical scale. Missing data were imputed using the K nearest neighbors' approach. Data were rescaled with min-max scaling.

Statistical Analysis

We developed 5 ML models using NAHR data: logistic regression, random forests,[27] neural networks,[28] gradient boosting trees,[29] and support vector machines (SVMs).[30] The patient cohort was split into 3; a training cohort, validation cohort, and test cohort with an 80:10:10 random split. The training cohort was used to train models and the validation cohort to tune the hyperparameters of each model. Ten-fold cross validation was used to reduce overfitting. Each method was optimized using a grid search over the hyperparameters to reach the greatest accuracy in the validation sets, before being deployed on the test set to yield final evaluation metrics. Feature importance was calculated using mean decrease in impurity, a metric that calculates the contribution of a particular feature to discriminating between classes within the random forest model.

In order to determine a minimum set of iHOT-12 questions required for accurate prediction of MCID, recursive feature elimination was used with our best-performing PROM-based model.[31] In this iterative process, the variable with the lowest feature importance calculated through mean decrease in impurity is removed with each run of the model to reduce the number of input variables to a core dataset. All statistical analysis was performed using Python (Python Software Foundation, Wilmington, DE).

Results

NAHR data for the study time period included 5,199 patients who underwent arthroscopy for FAIS. Of these patients, 1,917 had 6-month follow-up iHOT-12 overall scores recorded and were included in this study. The procedures performed on each patient are reported in Appendix Table 1, available at www.arthroscopyjournal.org.

Baseline Patient Characteristics

The included patient population was made up of 1,155 male and 762 female patients. Within the sample population, 1,266 of 1,917 (66.0%) achieved MCID. The characteristics of the sample population are detailed in Table 1, including the proportion of missing data for each baseline parameter. All patients had completed outcome data for the primary outcome, iHOT-12 overall score at 6 months' postoperatively.

Model Performance

The AUROC scores of the 5 models with all variables included are given in Table 2, and the AUROC curves for all models are shown in Figure 1A. Further metrics such as precision, recall, and F1 score are detailed in Appendix Table 2, available at www.arthroscopyjournal.org. With all variables included, the random forest model was the most accurate ML model, with fair predictive performance. The SVM and gradient boosting models also had fair predictive performance. The logistic regression and neural network models had poor performance. The performance of these models with validation data is presented in Appendix Table 3, available at www.arthroscopyjournal.org.

Preoperative Feature Importance by Category

The performance of the 5 model predictors using preoperative parameters is shown in Table 2. The AUROC curves are shown in Figure 1, B-G. Inclusion of either demographics or disease features alone resulted in decreased predictive performance across all models, as did inclusion of demographics and disease features together. Inclusion of all PROMs led to little change in AUROC compared with all variable random forest and gradient boosting models, slight increases in performance of logistic regression and neural network models, and no change in the SVM model. Inclusion of EQ-5D-5L PROMs alone as predictors resulted in a decrease in predictive performance across all models. Inclusion of iHOT-12 PROMs alone as predictors led to a decrease in model accuracy across all model types versus the inclusion of all variables. This decrease was marginal for random forest and neural network models.

Preoperative Feature Importance in Successful All Variable Models

The best-performing model, the random forest, was also used to isolate the features considered most important when making predictions with all variables and iHOT-12 PROMs alone used as predictive variables. Feature importance is displayed graphically in Figure 2, A and B. In the all-variable model features with high importance include EQ-5D-5L overall score, age, BMI and both iHOT-12 overall score and individual questions. In the iHOT-12 model feature importance of individual questions is similar.

iHOT-12 as a Minimum Dataset

To determine whether a specific combination of discrete iHOT-12 questions provided high performance

Table 1. Baseline Patient Data Presented for the Cohort of 1,917 Patients Identified From the Non-Arthroplasty Hip Registry Dataset

Characteristics	
Sex	
Male	1,155
Female	762
Age, y, n = 1,917 (100%)	35.9 ± 10.4
Laterality of operation	
Right hip	1,109
Left hip	808
Not recorded	10
BMI, n = 1,468 (76.6%)	25.7 ± 4.89
Outerbridge classification	
Grade 0	904
Grade 1	191
Grade 2	133
Grade 3	69
Grade 4	38
Not recorded	582
Surgery	
Primary	463
Revision	30
Second revision	1
Not recorded	1,423
Acetabular cartilage damage	
Cartilage damage present	1,255
Cartilage damage absent	218
Not recorded	444
Acetabular cartilage damage zone*	
Zone 1	111
Zone 2	879
Zone 3	604
Zone 4	103
Zone 5	20
Zone 6	15
Multiple zones	338
Not recorded	668
EQ-5D 5L: preoperative overall score, n = 1,915 (99.9%)	0.52 ± 0.23
EQ-5D 5L: Health VAS, n = 1,917 (100%)	67.4 ± 20.5
EQ-5D 5L: Preoperative mobility	
I have moderate problems in walking about	767
I have slight problems in walking about	580
I have no problems in walking about	300
I have severe problems in walking about	229
I am unable to walk about	20
Not recorded	21
EQ-5D 5L: Preoperative self-care	
I have no problems washing or dressing myself	1,181
I have slight problems washing or dressing myself	452
I have moderate problems washing or dressing myself	211
I have severe problems washing or dressing myself	43
I am unable to wash or dress myself	9
Not recorded	21
EQ-5D 5L: Preoperative usual activities	
I have moderate problems doing my usual activities	772
I have slight problems doing my usual activities	464
I have severe problems doing my usual activities	386
I have no problems doing my usual activities	143
I am unable to do my usual activities	131
Not recorded	21
EQ-5D 5L: Preoperative pain and discomfort	
I have moderate pain or discomfort	957
I have severe pain or discomfort	492
I have slight pain or discomfort	349
I have extreme pain or discomfort	69
I have no pain or discomfort	29
Not recorded	21
EQ-5D 5L: Preoperative anxiety and depression	
I am not anxious or depressed	869
I am slightly anxious or depressed	566
I am moderately anxious or depressed	352
I am severely anxious or depressed	78
I am extremely anxious or depressed	31
Not recorded	21
iHOT-12: Preoperative overall score, n = 1,917 (100%)	32.7 ± 17.9
iHOT-12 Q1: Preoperative level of pain, n = 1,828 (95.4%)	35.0 ± 20.9
iHOT-12 Q2: Difficulty getting up, n = 1,820 (94.9%)	45.0 ± 31.2
iHOT-12 Q3: Difficulty walking long distances, n = 1,825 (95.2%)	33.7 ± 27.8
iHOT-12 Q4: Trouble with grinding, catching, or clicking, n = 1,823 (95.1%)	36.0 ± 29.7
iHOT-12 Q5: Trouble with pushing, pulling, or lifting heavy objects, n = 1,810 (94.4%)	45.0 ± 30.6
iHOT-12 Q6: Concern regarding changing directions, n = 1,799 (93.8%)	27.6 ± 27.0
iHOT-12 Q7: Pain After Activity, n = 1,819 (94.9%)	22.7 ± 20.3
iHOT-12 Q8: Concern regarding carrying children, n = 1,774 (92.%)	47.7 ± 35.4
iHOT-12 Q9: Trouble with sexual activity, n = 1,625 (84.8%)	42.2 ± 30.8
iHOT-12 Q10: Time aware of disability, n = 1,823 (95.1%)	18.7 ± 20.8
iHOT-12 Q11: Concern regarding desired fitness level, n = 1,822 (95.0%)	15.4 ± 18.8
iHOT-12 Q12: Distraction attributable to hip problem, n = 1,823 (95.1%)	20.5 ± 19.0

NOTE. Parameters that are reported on a continuous scale, including VAS responses for PROMs, have been presented as mean ± standard deviation, number of patients with data for the given parameter (percentage response rate). Parameters that are reported in a categorical or ordinal manner have been presented with number of responses per category in the dataset and the number of missing entries.

BMI, body mass index; EQ-5D 5L, EuroQoL 5-Dimension 5-Level; iHOT-12, International Hip Outcome Tool; PROMs, patient-reported outcome measures; VAS, visual analog scale.

*Individual data points for acetabular cartilage damage zone are nonexclusive and individuals within the dataset who have damage across multiple zones have been recorded multiple times through zone 1-6.

in outcome prediction recursive feature elimination was used in our random forest model. The model was developed and refined across multiple runs with varied combinations of training and test data (Fig 3A), producing a mean output (Fig 3B). This shows that fair

Table 2. Performance of Machine-Learning Models on Test Set Data Indicated by Our Primary Outcome Measure, AUROC

	Performance (AUROC and 95% CI)						
	All Variables	Patient Demographics	Disease Features	Patient Demographics and Disease Features	All PROMs	EQ-5D 5L	iHOT-12
Logistic regression	0.69 (0.63-0.76)	0.55 (0.49-0.62)	0.46 (0.40-0.53)	0.53 (0.47-0.60)	0.70 (0.64-0.77)	0.52 (0.45-0.59)	0.60 (0.53-0.67)
Random forest	0.75 (0.68-0.81)	0.53 (0.46-0.60)	0.54 (0.47-0.61)	0.50 (0.43-0.57)	0.74 (0.68-0.80)	0.53 (0.46-0.60)	0.72 (0.65-0.78)
Neural network	0.69 (0.63-0.76)	0.55 (0.49-0.62)	0.49 (0.43-0.56)	0.47 (0.40-0.54)	0.72 (0.66-0.79)	0.49 (0.42-0.56)	0.68 (0.62-0.74)
Support vector machine	0.70 (0.64-0.77)	0.58 (0.51-0.65)	0.42 (0.36-0.49)	0.58 (0.51-0.65)	0.70 (0.64-0.77)	0.52 (0.45-0.59)	0.65 (0.59-0.72)
Gradient boosting	0.70 (0.64-0.77	0.50 (0.43-0.57)	0.51 (0.44-0.57)	0.50 (0.43-0.57)	0.69 (0.63-0.76)	0.54 (0.48-0.61)	0.64 (0.58-0.71)

NOTE. Each type of machine-learning model was produced using different combinations of preoperative predictors to ascertain relative input feature importance. Variation of input features is represented in each column.

AUROC, area under the receiver operating characteristic; CI, confidence interval; EQ-5D-5L, EuroQoL 5-Dimension 5-Level; iHOT-12, International Hip Outcome Tool.

predictive performance was maintained with a restricted dataset. Models consisting of 6 iHOT-12 questions achieved on average an AUROC of 0.71 ± 0.018 (Fig 3B). The iHOT-12 questions with the highest feature importance identified were as follows: "Pain," "Difficulty walking long distance," "Trouble with grinding," "Concern changing direction during sport," "Pain after activity," and "Trouble during sexual activity."

Discussion

This study used data from the UK's Non-Arthroplasty Hip Registry to show that ML techniques are able to predict with fair performance whether patients will achieve MCID of the iHOT-12 score after arthroscopic management of FAIS in a national cohort. Conclusions regarding the predictive ability of ML models reported here must be interpreted in light of wide AUROC confidence intervals and an analysis limited by a high proportion of missing data. This study also sought to determine which features are important for ML-based outcome prediction and to generate predictive ML models with fewer preoperative features. When preoperative predictor classes were used as predictive features, both demographics and disease features had poor predictive ability, whereas PROMs and iHOT-12 had fair predictive ability. The responses to the iHOT-12 questionnaire alone were sufficient to predict MCID attainment with similar performance to entire dataset models. This predictive ability was retained using just 6 question responses from the iHOT-12.

The primary aim of this study was to use NAHR data to determine whether ML techniques could predict accurately whether patients would achieve MCID on the iHOT-12 score after arthroscopic surgery for FAIS. The performance of models used in this study enabled fair predictive performance. The random forest model achieved fair performance, with an AUROC of 0.75 and an F1 score of 0.84. A previous study achieved good predictive performance using logistic regression-based methods in order to predict MCID after FAIS in the Hip Outcome Score and modified Harris Hip Score PROMs.[17] In another study, Kunze et al.[18] achieved good predictive performance in the prediction of Hip Outcome Score-Activities of Daily Living scores with stochastic gradient boosting approaches with a final model AUROC of 0.84. Comparatively, the model presented in this work has inferior predictive performance. Both of these studies used a single-surgeon dataset for analysis. This has implications regarding external validity for the models developed and their real-world utility. In addition, these datasets had a larger variety of features with a generally lower proportion of missing data in each feature, this, in combination with a more homogenous cohort enabling greater accuracy of prediction. For ML models to be useful tools in health care,

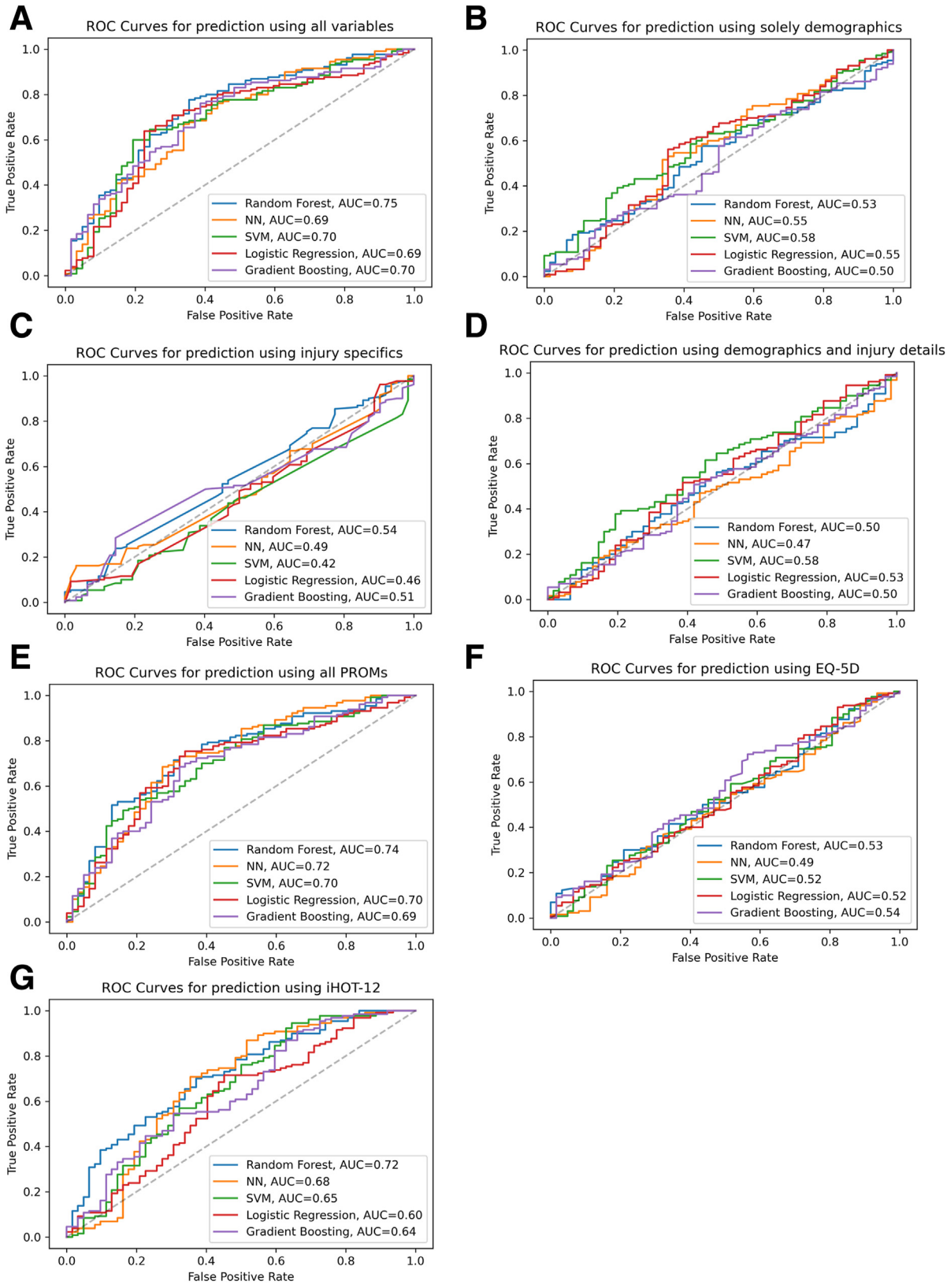

Fig 1. Receiver operator curves for model performance on test data, with AUC indicating the area under the receiver operator curve (AUROC). Random forest performance is indicated in blue, neural network in yellow, support vector machine in green, logistic regression in red, and gradient boosting models in purple. Predictive features used vary between image: (A) All variables, (B) demographics alone, (C) disease features, (D) both demographics and disease features, (E) all baseline PROMs, (F) baseline EQ-5D-5L alone, and (G) baseline iHOT-12 alone. (EQ-5D-5L, EuroQoL 5-Dimension 5-Level; iHOT-12, International Hip Outcome Tool; PROMS, patient-reported outcome measures.)

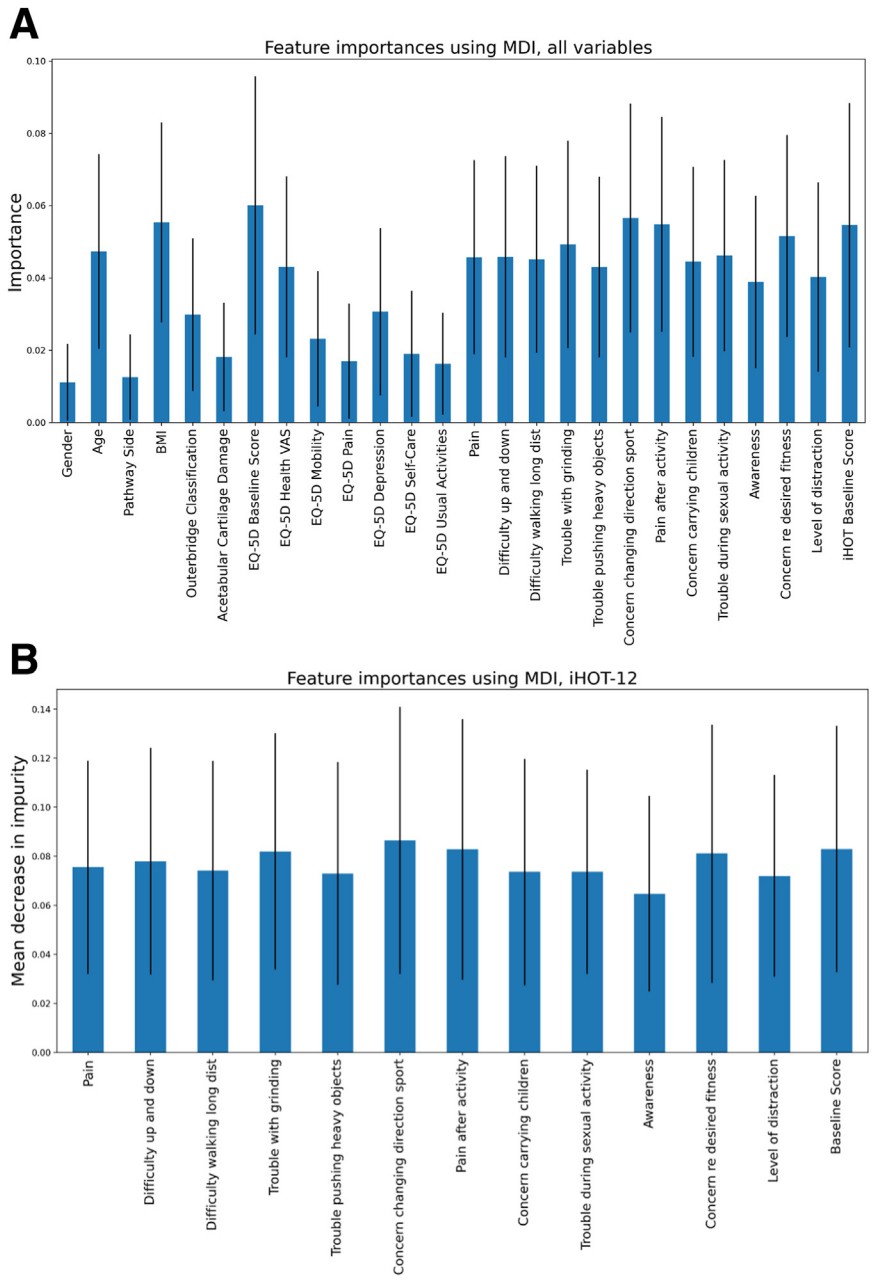

Fig 2. Graph plotting individual predictive feature importance in the best-performing predictive model, the random forest model, with input features as: (A) all variables and (B) baseline iHOT-12 questions alone. (iHOT-12, International Hip Outcome Tool.)

they require development using diverse data across large numbers of patients throughout an entire population; however, the data must be of a sufficient quality with full data collection and appropriate follow-up of patients. Both of these caveats are issues in the use of registry data, which inherently have variable quality data submission and have little active patient follow-up.[32]

This study presents an ML model for hip arthroscopy MCID attainment prediction developed using national registry data, predicting MCID attainment in the iHOT-12 following hip arthroscopy. In the NAHR sample data for this study, just fewer than 37% of patients who had undergone arthroscopy for FAIS had completed the defined primary outcome, limiting the dataset available for model development, and of this 37%, many patients had missing data for baseline patient variables as detailed in Table 1. Furthermore, there was expected variance in the dataset due to the multisurgeon and multicenter nature of the data. The authors believe that both of these features contribute to the lower predictive performance achieved with this study versus that of Kunze et al.[18]

Similar findings have been published by Martin et al.,[33] who used Danish national registry data to predict progression to revision arthroscopy after primary hip arthroscopy. Martin et al.[33] achieved comparatively poor performance to ML models

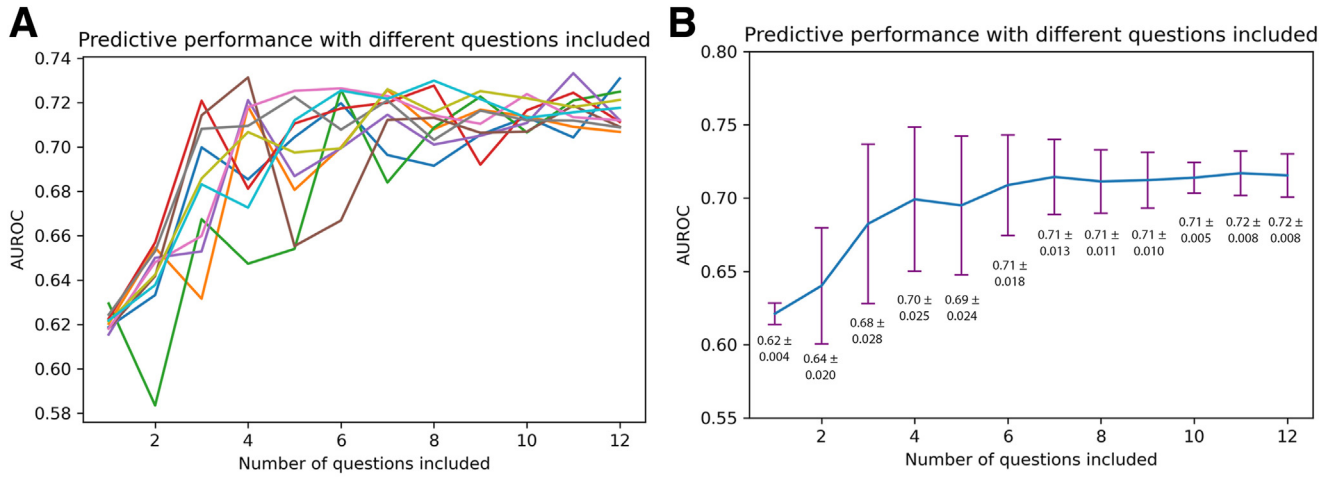

Fig 3. Graph plotting AUROC achieved versus number of baseline iHOT-12 question responses used as predictive features by the random forest model during the process of recursive feature elimination for (A) varied combinations of training and test data; therefore, order of feature dropping and (B) an averaged model with mean ± standard deviation. (AUROC, area under the receiver operator curve; iHOT-12, International Hip Outcome Tool.)

developed using single-surgeon datasets, and concluded their models are of limited clinical utility due to wide discrimination confidence intervals. Results in this study have the same limitation; the AUROC for the best-performing model was 0.74 (0.68-0.81), which crosses into the category of poor performance. In addition, comparisons between model performance presented in Table 1 by ML methodology are limited due to overlapping confidence intervals shared by ML models. The authors believe the close grouping of model performance is a consequence of generally wide confidence intervals and poor predictive performance across all 5 ML methods resulting from training data quality, rather than there being no difference in ML model predictive utility given an ideal dataset. The results from this study are therefore limited by the quantity and quality of diverse data within the NAHR at this point. In the future, with improved data quality and improved representation of varied surgeons and institutions, further ML analysis may achieve improved predictive performance.

The secondary aim of this study was to determine which preoperative factors contribute to the predictive performance of these ML models and to achieve a reduced set of baseline features required for accurate outcome prediction. When comparing preoperative feature importance by category, across all models there was a decrease in performance using either patient demographics or disease features alone, or demographics and disease features in combination, compared with all variables. Particularly in cases in which the input variables were limited to just injury specifics or demographics, many models reverted to predicting that all patients reached MCID. This is exemplified by the precision and recall data shown in Appendix Table 2. These models tend to achieve a precision of, or approaching 0.66, and a recall of 1. These models simply predict that all patients attain MCID, and therefore generate true positives in 66% of cases with 34% false positives, providing a precision of 0.66, the proportion of patients who achieved MCID in our dataset. In addition, these models generate no false negatives, as they predict all patients achieve MCID, resulting in a recall of 1. Ultimately, this reflects that these predictor classes within the NAHR dataset contain too little information to enable the generation of effective ML models. In contrast, there was a small decrease, equal, or improved performance with models developed using all PROMs as input features. This suggests that PROMs alone contain sufficient information to adequately predict which patients will achieve MCID in the context of our dataset. Interestingly, when using just EQ-5D-5L as a predictor, models lose predictive power. This may be a result of a lower pathology specificity of the EQ-5D-5L, which is a more generalized health questionnaire. When using iHOT-12 scores alone, however, models perform well predictively and have very similar performance to all PROM models.

Our study identified individual features which contributed with high and low importance to the predictive power of our most successful model, the random forest. The feature with the highest importance appears to be overall EQ-5D-5L score, whereas other important features include age, BMI, and iHOT-12 scores. Recent systematic reviews have highlighted predictors of outcome for FAIS after arthroscopy. The following factors were deemed to be important: age, sex, BMI, Tönnis grade, chondral defects, decreased joint space ($\leq$2 mm), increased Kellgren–Lawrence grade (>3), increased lateral center-edge angle, and relief from

preoperative intra-articular corticosteroid injections.[34-36] In contrast to previous studies, our model ranked sex and acetabular cartilage damage as being relatively unimportant in predicting outcomes. The relatively low importance of these predictor classes in our study also may be explained by a sparsity of data requiring imputation and the low number of input variables across these classes in our data set. In addition, the high feature importance of overall PROM scores in EQ-5D-5L and iHOT-12 may indicate the patient's baseline perception of their performance state is a highly important predictor of outcome. Further work using ML models with more diverse datasets is required to determine whether all of the preoperative features used here are strongly associated with outcome, and, to determine which further preoperative features may be useful to collect in registries ongoing to maximize the potential of ML-based prediction tools. In addition, although not enabling entirely preoperative prediction of outcome, the inclusion of perioperative data such as the magnitude of bony corrections and the types of labral and regenerative procedures performed may enable further refinement of ML models and inform surgical planning.

The random forest model using iHOT-12 questions alone as predictive features displayed comparable performance to the model with all variable inputs. In analysis of feature performance, each question within the iHOT-12 has a similar importance. This suggests the overall impression of hip-specific health gained from the iHOT-12 questionnaire is important rather than the discerning nature of any individual question. This may be the result of respondent fatigue, with similar answers given to each question, or a case in which fewer questions still are required to account for variance in the iHOT-12 scores.[37] iHOT-12 questions account for 96% of the variation in iHOT-33 scores.[38] In further analysis using recursive feature elimination fair performance was achieved consistently until less than 6 iHOT-12 questions were used as input variables. Whilst recursive feature elimination identified questions which consistently displayed higher feature importance than other variables, this process displays that only 6 of the 12 iHOT-12 questions are required, on average, to predict MCID attainment with similar performance to whole data set models. The clinical utility of this finding is, however, limited by the wide confidence intervals and overall performance of whole dataset ML models produced in this work and as such the authors cannot recommend further exploration of this in clinical populations. It may be interesting to observe whether similar ML based explorations of outcome find that models using limited PROM instruments are able to provide similar performance to those including full PROM instruments in the future.

Ultimately the authors suggest that for accurate ML based tools for outcome prediction with high real-world utility, improved registry datasets will be required for robust ML model development.

Limitations

Whilst this study seeks to apply ML approaches to predicting iHOT-12 MCID attainment in patients with FAIS after hip arthroscopy and to predict MCID attainment after hip arthroscopy using national registry data, there are several limitations that have been highlighted in this discussion. The nature of the NAHR dataset in its current format is not optimized for ML algorithm development. There is a limited quantity of patients with data for the chosen primary outcome and the data quality is reduced by missing data regarding baseline patient features. In addition, due to the nature of the registry and relatively low number of patients appropriate for inclusion over a 9-year period, there will be high variability between surgeon and center reporting. Some surgeons may be overrepresented within the database at present, resulting in selection bias. A number of patient-specific variables have not been included in this analysis, which may be necessary to increase predictive success and for pragmatic considerations regarding surgical candidacy. Our inclusion and exclusion criteria were unable to take into account patients with comorbid hip pathology or complex surgical backgrounds, reducing performance especially in outlier cases. Finally, to enable the most accurate prediction of surgical outcome, a model should include perioperative factors, as tools based entirely on preoperative factors, the aforementioned models, are limited.

Conclusions

ML models were able to predict with fair accuracy, attainment of MCID on the iHOT-12 at the 6-month postoperative assessment. The most successful models used all patient variables, all baseline PROMs, and baseline iHOT-12 responses. These models are not sufficiently accurate to warrant routine use in the clinic currently.

Disclosure

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: A.M. reports grants, personal fees, and other from Pfizer, and grants from Schuelke and Vyfor, outside the submitted work. V.K. reports personal fees from Smith & Nephew and Arthrex, outside the submitted work. In addition, V.K. has a patent, Pressure Sensor – ArtioSense pending to ArtioSense. All other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Full ICMJE author

disclosure forms are available for this article online, as supplementary material.

References

1. Doran C, Pettit M, Singh Y, Sunil Kumar KH, Khanduja V. Does the type of sport influence morphology of the hip? A systematic review. *Am J Sports Med* 2022;50:1727-1741.
2. Pettit M, Doran C, Singh Y, Saito M, Sunil Kumar KH, Khanduja V. How does the cam morphology develop in athletes? A systematic review and meta-analysis. *Osteoarthr Cartil* 2021;29:1117-1129.
3. Ganz R, Parvizi J, Beck M, Leunig M, Nötzli H, Siebenrock KA. Femoroacetabular impingement: A cause for osteoarthritis of the hip. *Clin Orthop Relat Res* 2003;417:112-120.
4. Hale RF, Melugin HP, Zhou J, et al. Incidence of femoroacetabular impingement and surgical management trends over time. *Am J Sports Med* 2021;49:35-41.
5. Palmer AJR, Malak TT, Broomfield J, et al. Past and projected temporal trends in arthroscopic hip surgery in England between 2002 and 2013. *BMJ Open Sport Exerc Med* 2016;2(1).
6. Addai D, Zarkos J, Pettit M, Sunil Kumar KH, Khanduja V. Outcomes following surgical management of femoroacetabular impingement: A systematic review and meta-analysis of different surgical techniques. *Bone Jt Res* 2021;10:574-590.
7. Lu V, Andronic O, Zhang JZ, Khanduja V. Outcomes of arthroscopy of the hip for femoroacetabular impingement based on intraoperative assessment using the Outerbridge classification. *Bone Joint J* 2023;105:751-759.
8. Holleyman R, Sohatee MA, Lyman S, Malviya A, Khanduja V, NAHR User Group. Hip arthroscopy for femoroacetabular impingement is associated with significant improvement in early patient reported outcomes: analysis of 4963 cases from the UK non-arthroplasty registry (NAHR) dataset. *Knee Surg Sports Traumatol Arthrosc* 2023;31:58-69.
9. Van Houcke J, Khanduja V, Audenaert EA. Accurate arthroscopic cam resection normalizes contact stresses in patients with femoroacetabular impingement. *Am J Sports Med* 2021;49:42-48.
10. Palmer AJR, Ayyar Gupta V, Fernquest S, et al; FAIT Study Group. Arthroscopic hip surgery compared with physiotherapy and activity modification for the treatment of symptomatic femoroacetabular impingement: multicentre randomised controlled trial. *BMJ* 2019;364:l185. doi:10.1136/bmj.l185. Erratum in: *BMJ*. 2021 Jan 18;372:m3715.
11. Rosinsky PJ, Chen JW, Yelton MJ, et al. Does failure to meet threshold scores for mHHS and iHOT-12 correlate to secondary operations following hip arthroscopy? *J Hip Preserv Surg* 2020;7:272-280.
12. Mas Martinez J, Bustamante Suarez de Puga D, Verdu-Roman C, Martinez Gimenez E, Morales Santias M, Sanz-Reig J. Significant improvement after hip arthroscopy for femoroacetabular impingement in women. *Knee Surg Sport Traumatol Arthrosc* 2022;30:2181-2187.
13. Parvaresh K, Rasio JP, Martin RRL, et al. Achievement of meaningful clinical outcomes is unaffected by capsulotomy type during arthroscopic treatment of femoroacetabular impingement syndrome: Results from the Multicenter Arthroscopic Study of the Hip (MASH) Study Group. *Am J Sports Med* 2021;49:713-720.
14. Mas Martinez J, Verdu-Roman C, Bustamante Suarez de Puga D, Morales Santias M, Martinez Gimenez E, Sanz-Reig J. Arthroscopic surgery for femoroacetabular impingement has limited effect in patients with Tönnis grade-2 at 4-year follow-up. *Arch Orthop Trauma Surg* 2022;142:2801-2809.
15. Punnoose A, Claydon-Mueller LS, Weiss O, Zhang J, Rushton A, Khanduja V. Prehabilitation for patients undergoing orthopedic surgery: A systematic review and meta-analysis. *JAMA Netw Open* 2023;6:e238050.
16. Harris JD. Editorial Commentary: Personalized hip arthroscopy outcome prediction using machine learning-the future is here. *Arthroscopy* 2021;37:1498-1502.
17. Nwachukwu BU, Beck EC, Lee EK, et al. Application of machine learning for predicting clinically meaningful outcome after arthroscopic femoroacetabular impingement surgery. *Am J Sports Med* 2020;48:415-423.
18. Kunze KN, Polce EM, Nwachukwu BU, Chahla J, Nho SJ. Development and internal validation of supervised machine learning algorithms for predicting clinically significant functional improvement in a mixed population of primary hip arthroscopy. *Arthroscopy* 2021;37:1488-1497.
19. Kunze KN, Polce EM, Clapp IM, Alter T, Nho SJ. Association between preoperative patient factors and clinically meaningful outcomes after hip arthroscopy for femoroacetabular impingement syndrome: A machine learning analysis. *Am J Sports Med* 2022;50:746-756.
20. Wellington IJ, Cote MP. Editorial Commentary: Machine learning in orthopaedics: Venturing into the valley of despair. *Arthroscopy* 2022;38:2767-2768.
21. Pareek A, Martin RK. Editorial Commentary: Machine learning in medicine requires clinician input, faces barriers, and high-quality evidence is required to demonstrate improved patient outcomes. *Arthroscopy* 2022;38:2106-2108.
22. Kunze KN, Krivicich LM, Clapp IM, et al. Machine learning algorithms predict achievement of clinically significant outcomes after orthopaedic surgery: A systematic review. *Arthroscopy* 2022;38:2090-2105.
23. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med* 2015;162:W1-W73.
24. Nwachukwu BU, Chang B, Beck EC, et al. How should we define clinically significant outcome improvement on the iHOT-12? *HSS J* 2019;15:103-108.
25. Martin RL, Kivlan BR, Christoforetti JJ, et al. Minimal clinically important difference and substantial clinical benefit values for the 12-Item International Hip Outcome Tool. *Arthroscopy* 2019;35:411-416.
26. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;240:1285-1293.
27. Breiman L. Random forests. *Mach Learn* 2001;45:5-32.
28. Hecht-Nielsen R. Theory of the backpropagation neural network In: *International 1989 Joint Conference on Neural Networks*1;593-605. doi:10.1109/IJCNN.1989.118638

29. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat* 2001;29:1189-1232.
30. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20:273-279.
31. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn* 2002;46:389-422.
32. Rubinger L, Ekhtiari S, Gazendam A, Bhandari M. Registries: Big data, bigger problems? *Injury* 2023;54:S39-S42 (suppl 3).
33. Martin RK, Wastvedt S, Lange J, Pareek A, Wolfson J, Lund B. Limited clinical utility of a machine learning revision prediction model based on a national hip arthroscopy registry. *Knee Surg Sport Traumatol Arthrosc* 2023;31:2079-2089.
34. Kuroda Y, Saito M, Çınar EN, Norrish A, Khanduja V. Patient-related risk factors associated with less favourable outcomes following hip arthroscopy. *Bone Joint J* 2020;102:822-831.
35. Sogbein OA, Shah A, Kay J, et al. Predictors of outcomes after hip arthroscopic surgery for femoroacetabular impingement: A systematic review. *Orthop J Sport Med* 2019;7:1-19.
36. Kuroda Y, Saito M, Çınar EN, Norrish A, Khanduja V. Patient-related risk factors associated with less favourable outcomes following hip arthroscopy. *Bone Joint J* 2020;102-B:822-831.
37. Rolstad S, Adler J, Rydén A. Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value Heal* 2011;14:1101-1108.
38. Griffin DR, Parsons N, Mohtadi NGH, Safran MR. A short version of the International Hip Outcome Tool (iHOT-12) for use in routine clinical practice. *Arthroscopy* 2012;28:611-618.

Appendix Tables

Appendix Table 1. Arthroscopic Procedures Performed for the Cohort of 1,917 Patients Identified From the Non-Arthroplasty Hip Registry Dataset

Acetabular procedure	
Acetabular labral debridement	926
Acetabular labral repair	890
Acetabular rim recession	659
Acetabular microfracture	118
Subspinous resection	38
Autologous chondrocyte implantation	23
Other	73
Femoral procedure	
Cam removal	1,644
Femoral osteophyte removal	56
Femoral cartilage debridement	23
Femoral microfracture	9
Other	7
Other procedure	
Psoas release	53
Ligamentum teres debridement	30
Loose body removal	20
Trochanteric bursa debridement	12
ITB release	4
Ligamentum teres reconstruction	1

ITB, iliotibial band.

Appendix Table 2. Performance of Machine-Learning Models on Test Set Data Presented With Variation of Input Features by the Preoperative Predictor

		Performance (F1 Score, Precision, and Recall)						
		All Variables	Patient Demographics	Disease Features	Patient Demographics and Disease Features	All PROMs	EQ-5D	iHOT-12
Logistic regression	F1 Score	0.78	0.80	0.80	0.80	0.76	0.80	0.82
	Precision	0.69	0.66	0.66	0.66	0.67	0.68	0.70
	Recall	0.89	1.00	1.00	1.00	0.94	0.94	0.97
Random forest	F1 Score	0.84	0.79	0.80	0.81	0.82	0.81	0.84
	Precision	0.73	0.67	0.66	0.68	0.71	0.69	0.72
	Recall	0.99	0.95	1.00	0.99	0.98	0.97	1.00
Neural network	F1 Score	0.82	0.80	0.80	0.80	0.82	0.80	0.82
	Precision	0.75	0.66	0.66	0.66	0.70	0.66	0.74
	Recall	0.90	1.00	1.00	1.00	1.00	1.00	0.91
Support vector machine	F1 Score	0.81	0.80	0.80	0.80	0.79	0.80	0.82
	Precision	0.69	0.66	0.66	0.66	0.68	0.66	0.69
	Recall	0.98	1.00	1.00	1.00	0.95	1.00	0.99
Gradient boosting	F1 Score	0.80	0.75	0.80	0.78	0.81	0.78	0.82
	Precision	0.72	0.66	0.67	0.66	0.72	0.69	0.71
	Recall	0.92	0.86	0.98	0.95	0.95	0.88	0.95

NOTE. Performance indicated by our secondary outcome measures F1 score, precision, and recall.
EQ-5D 5L, EuroQoL 5-Dimension 5-Level; iHOT-12, International Hip Outcome Tool; PROM, patient-reported outcome measure.

Appendix Table 3. Performance of Machine-Learning Models on Validation Data Indicated by Our Primary Outcome Measure, AUROC

	Performance (AUROC and 95% CI)						
	All Variables	Patient Demographics	Disease Features	Patient Demographics and Disease Features	All PROMs	EQ-5D	iHOT-12
Logistic regression	0.70 (0.63-0.76)	0.55 (0.49-0.62)	0.47 (0.41-0.54)	0.53 (0.46-0.59)	0.71 (0.65-0.78)	0.52 (0.46-0.59)	0.60 (0.53-0.67)
Random forest	0.76 (0.69-0.82)	0.54 (0.47-0.60)	0.54 (0.47-0.61)	0.52 (0.45-0.58)	0.74 (0.68-0.80)	0.54 (0.47-0.60)	0.74 (0.67-0.80)
Neural network	0.70 (0.63-0.76)	0.56 (0.50-0.63)	0.50 (0.44-0.57)	0.47 (0.41-0.54)	0.73 (0.67-0.80)	0.49 (0.42-0.56)	0.69 (0.62-0.75)
Support vector machine	0.70 (0.63-0.76)	0.60 (0.53-0.67)	0.49 (0.43-0.56)	0.60 (0.53-0.66)	0.71 (0.64-0.78)	0.53 (0.46-0.60)	0.65 (0.59-0.72)
Gradient boosting	0.70 (0.64-0.77	0.50 (0.44-0.57)	0.51 (0.45-0.58)	0.52 (0.45-0.58)	0.70 (0.64-0.77)	0.54 (0.47-0.60)	0.65 (0.59-0.71)

NOTE. Each type of machine-learning model was produced using different combinations of pre-operative predictors to ascertain relative input feature importance. Variation of input features is represented in each column.

AUROC, area under the receiver operator curve; CI, confidence interval; EQ-5D 5L, EuroQoL 5-Dimension 5-Level; iHOT-12, International Hip Outcome Tool; PROM, patient-reported outcome measure.