Using Machine Learning to Predict Clinical Outcomes After Shoulder Arthroplasty with a Minimal Feature Set

Article *in* Journal of shoulder and elbow surgery / American Shoulder and Elbow Surgeons ... [et al.] · August 2020

DOI: 10.1016/j.jse.2020.07.042

CITATIONS
4

READS
155

9 authors, including:

Vikas Kumar
Chandigarh University
28 PUBLICATIONS **173** CITATIONS
SEE PROFILE

Chris Roche
Exactech, Inc.
129 PUBLICATIONS **1,861** CITATIONS
SEE PROFILE

Ryan Simovitch
Palm Beach Orthopaedic Institute
35 PUBLICATIONS **1,466** CITATIONS
SEE PROFILE

J Shoulder Elbow Surg (2020) ■, 1–12

ELSEVIER

Journal of Shoulder and Elbow Surgery

www.elsevier.com/locate/ymse

Using machine learning to predict clinical outcomes after shoulder arthroplasty with a minimal feature set

Vikas Kumar, PhD[a], Christopher Roche, MSE, MBA[b,*], Steven Overman, MD, MPH[a], Ryan Simovitch, MD[c], Pierre-Henri Flurin, MD[d], Thomas Wright, MD[e], Joseph Zuckerman, MD[f], Howard Routman, DO[g], Ankur Teredesai, PhD[a]

[a]*KenSci, Seattle, WA, USA*
[b]*Exactech, Gainesville, FL, USA*
[c]*Hospital for Special Surgery Florida, West Palm Beach, FL, USA*
[d]*Bordeaux-Merignac Sport Clinic, Merignac, France*
[e]*Department of Orthopaedic Surgery, University of Florida, Gainesville, FL, USA*
[f]*Department of Orthopedic Surgery, NYU Langone Orthopedic Hospital, New York, NY, USA*
[g]*Atlantis Orthopedics, Palm Beach Gardens, FL, USA*

Background: A machine learning analysis was conducted on 5774 shoulder arthroplasty patients to create predictive models for multiple clinical outcome measures after anatomic total shoulder arthroplasty (aTSA) and reverse total shoulder arthroplasty (rTSA). The goal of this study was to compare the accuracy associated with a full–feature set predictive model (ie, full model, comprising 291 parameters) and a minimal–feature set model (ie, abbreviated model, comprising 19 input parameters) to predict clinical outcomes to assess the efficacy of using a minimal feature set of inputs as a shoulder arthroplasty clinical decision-support tool.

Methods: Clinical data from 2153 primary aTSA patients and 3621 primary rTSA patients were analyzed using the XGBoost machine learning technique to create and test predictive models for multiple outcome measures at different postoperative time points via the full and abbreviated models. Mean absolute errors (MAEs) quantified the difference between actual and predicted outcomes, and each model also predicted whether a patient would experience clinical improvement greater than the patient satisfaction anchor-based thresholds of the minimal clinically important difference and substantial clinical benefit for each outcome measure at 2-3 years after surgery.

Results: Across all postoperative time points analyzed, the full and abbreviated models had similar MAEs for the American Shoulder and Elbow Surgeons score (±11.7 with full model vs. ±12.0 with abbreviated model), Constant score (±8.9 vs. ±9.8), Global Shoulder Function score (±1.4 vs. ±1.5), visual analog scale pain score (±1.3 vs. ±1.4), active abduction (±20.4° vs. ±21.8°), forward elevation (±17.6° vs. ±19.2°), and external rotation (±12.2° vs. ±12.6°). Marginal improvements in MAEs were observed for each outcome measure prediction when the abbreviated model was supplemented with data on implant size and/or type and measurements of native glenoid anatomy. The full and abbreviated models each effectively risk stratified patients using only preoperative data by accurately identifying patients with improvement greater than the minimal clinically important difference and substantial clinical benefit thresholds.

Discussion: Our study showed that the full and abbreviated machine learning models achieved similar accuracy in predicting clinical outcomes after aTSA and rTSA at multiple postoperative time points. These promising results demonstrate an efficient utilization of machine learning algorithms to predict clinical outcomes. Our findings using a minimal feature set of only 19 preoperative inputs suggest that this tool may be easily used during a surgical consultation to improve decision making related to shoulder arthroplasty.

All data were acquired in an institutional review board (IRB)–approved study (NYU IRB study no. i05-144_MOD41) that was carried out in accordance with relevant regulations of the US Health Insurance Portability and Accountability Act.

*Reprint requests: Christopher Roche, MSE, MBA, Exactech, 2320 NW 66th Ct, Gainesville, FL 32653, USA.
E-mail address: Chris.Roche@exac.com (C. Roche).

https://doi.org/10.1016/j.jse.2020.07.042

Level of evidence: Basic Science Study; Computer Modeling

Keywords: Machine learning; predictive outcomes analytics; aTSA outcomes; rTSA outcomes; shoulder arthroplasty; clinical research

Machine learning techniques can analyze clinical and patient-reported outcome data to create predictive models that can help physicians better understand their patients prior to treatment by quantifying patient-specific potential for improvement associated with different treatment options. The knowledge of these patient-specific outcome predictions is useful to better inform shared decision making for both the patient and the surgeon.[2-4,12] Machine learning models have recently been used to accurately predict clinical outcomes after anatomic total shoulder arthroplasty (aTSA) and reverse total shoulder arthroplasty (rTSA) and to risk stratify patients based on predicted minimal clinically important difference (MCID) and substantial clinical benefit (SCB) improvement thresholds for different clinical outcome measures.[12] Deploying such a preoperative prediction tool in clinical practice offers the potential to establish more accurate expectations of patient-specific improvement that can be achieved with shoulder arthroplasty, as well as to align the patient and surgeon on what results to expect at different postoperative time points. The practical limitation of deploying such a tool in the clinic is the large input burden often required by machine learning algorithms to generate patient-specific predictions, particularly because much of this information may not be routinely present in the patient's electronic medical record. As such, a prerequisite for a machine learning–based clinical decision-support tool is the identification of a highly predictive minimal set of preoperative inputs that can be readily obtained as part of the normal standard of care.

To develop a clinical decision-support tool using a minimal feature set of preoperative inputs, we first conducted a machine learning analysis on a multicenter clinical database of 1 platform shoulder prosthesis to create algorithms using the full set of preoperative data to predict postoperative outcomes of various measures at multiple postoperative time points after aTSA and rTSA. We developed these algorithms using a full feature set (ie, full model, comprising 291 input parameters), and in the process of doing so, we identified a minimal feature set (ie, abbreviated model, comprising 19 input parameters) consisting only of the most highly predictive features. Therefore, the goal of this study was to quantify and compare the accuracy associated with the full and abbreviated machine learning models to predict clinical outcomes after aTSA and rTSA.

Methods

We used the XGBoost machine learning technique[22] to analyze a multicenter clinical outcomes database of shoulder arthroplasty patients who received a single platform shoulder prosthesis (Equinoxe; Exactech, Gainesville, FL, USA) between November 2004 and December 2018. Every patient enrolled in this open-label clinical database provided consent. All data were collected using standardized forms at each of the 30 clinical sites. On completion of each form, each is independently verified and then scored by a computer on a secured IBM database (IBM, Armonk, NY, USA). All primary aTSA and primary rTSA patients in the database with $\geq$3 months of follow-up were included. To ensure a homogeneous data set, patients with revisions, humeral fractures, endoprostheses, and hemiarthroplasty were excluded. Primary total shoulder arthroplasty patients who experienced complications and/or revisions were not excluded from this analysis as those experiences contribute to the outcome variability of the aTSA and rTSA cohorts.

The aforementioned criteria resulted in preoperative, intraoperative, and postoperative data from 5774 patients with 17,427 postoperative follow-up visits available for analysis in this machine learning study. The full database contains 291 inputs, including demographic characteristics, diagnoses, comorbidities, implant type, range of motion (ROM), radiographic findings, and clinical outcome metric scores (American Shoulder and Elbow Surgeons [ASES], Constant, University of California–Los Angeles, Simple Shoulder Test, and Shoulder Pain and Disability Index), as well as the individual questions used to derive each of these patient-reported outcome scores. ROM assessment was performed by the implanting surgeon or this surgeon's surrogate and was measured with a goniometer. These data were used to create predictive algorithms for the ASES score, Constant score, Global Shoulder Function score, visual analog scale (VAS) pain score, active abduction, active forward elevation, and active external rotation with the arm at the side at multiple time points after aTSA or rTSA, including 3-6 months, 6-9 months, 1 year (9-18 months), 2-3 years (18-36 months), 3-5 years (36-60 months), and $\geq$5 years ($\geq$60 months).

XGBoost was used to create the predictive algorithms for both the full and abbreviated models. XGBoost is a supervised, ensemble machine learning technique of multiple-regression trees that are built by iteratively partitioning the training data set into multiple small batches using a method called "boosting."[22] The full model used all 291 inputs from the database, whereas the abbreviated model used only a minimal feature set of 19 preoperative inputs (Table I) to predict the Global Shoulder Function score; VAS pain score; and active abduction, forward elevation, and external rotation. As described in Table I, this minimal feature set is a selection of patient demographic characteristics, diagnoses, comorbidities, preoperative ROM, and patient responses to a few subjective questions. These specific 19 input parameters were identified using domain knowledge, the prevalence of the feature, the uniqueness of the feature values for patients, and finally, the importance of features to the model. Of note, the uniqueness of a feature is computed using an information theory metric known as "entropy," which measures whether values in a

Table I Description of minimal feature set of preoperative inputs utilized by abbreviated prediction model

Feature	Description and unit	Range or inputs
Age	Age of patient, yr	18-115
Weight	Weight, lb	80-450 lb (36.3-204.1 kg)
Height	Height, in	48-80 in (121.9-203.2 cm)
Sex	Male or female	Male or female
Previous shoulder surgery	Has the patient previously had a surgical operation on the shoulder?	Yes or no
Surgery on dominant shoulder	Will the upcoming arthroplasty be on the patient's dominant shoulder?	Yes or no
Primary diagnosis	What is the patient's diagnosis?	Osteoarthritis, osteonecrosis, rotator cuff tear, rotator cuff tear arthropathy, rheumatoid arthritis, or post-traumatic arthritis
Comorbidities	What are the patient's comorbidities?	No comorbidities, inflammatory arthritis, hypertension, heart disease, diabetes, chronic renal failure, or tobacco use
Preoperative active abduction	Active arm elevation in frontal plane, °	0-180°
Preoperative active forward elevation	Active arm elevation in sagittal plane, °	0-180°
Preoperative active external rotation	Active lateral rotation of arm, with arm at side, °	−90 to 90°
Preoperative passive external rotation	Passive lateral rotation of arm, with arm at side, °	−90 to 90°
Preoperative internal rotation score	Active medial rotation of arm, with arm at side; unitless: 8-point numeric scale with the following discreet assignments based on motion to vertebral segments: 0, no motion; 1, hip; 2, buttocks; 3, sacrum; 4, L5 to L4; 5, L3 to L1; 6, T12 to T8; and 7, T7 or higher	0-7
Preoperative Global Shoulder Function score	Patient assessment of ability to use shoulder prior to surgery via Global Shoulder Function score; 11-point score (0-10), with 10 indicating full or normal mobility	0-10
Preoperative pain on daily basis (ie, VAS score)	Patient assessment of pain experienced on daily basis prior to surgery via VAS pain score; 11-point score (0-10), with 10 indicating extreme pain	0-10
Preoperative pain at worst	Patient assessment of worst pain experienced on daily basis prior to surgery; 11-point score (0-10), with 10 indicating extreme pain	0-10
Preoperative pain when lying on side	Patient assessment of pain experienced when lying on affected side prior to surgery; 11-point score (0-10), with 10 indicating extreme pain	0-10
Preoperative pain when touching back of neck	Patient assessment of pain experienced when touching back of neck prior to surgery; 11-point score (0-10), with 10 indicating extreme pain	0-10
Preoperative pain when pushing with affected arm	Patient assessment of pain experienced when pushing with affected arm prior to surgery; 11-point score (0-10), with 10 indicating extreme pain	0-10

VAS, visual analog scale.

feature are highly uncertain and thus likely to be highly random across patients, making it difficult to predict outcomes based on that feature. Moreover, the importance of a feature to the model was computed based on F-scores determined from the XGBoost algorithm. The F-score quantifies the frequency at which a particular feature is used as a candidate for the split by the decision-tree algorithm.[22] The prevalence, entropy, and F-score were used to determine an individual ranking of each feature by combining them into a single ranked list using the reciprocal fusion rank score. When this abbreviated model is supplemented with the 10 additional questions needed to calculate the preoperative ASES score and the 20 additional questions needed to calculate the preoperative Constant score, the abbreviated model can also be used to predict the ASES and Constant scores at each postoperative time point for both aTSA and rTSA. Finally, we conducted an additional analysis that supplemented the

abbreviated model predictions with data on implant size and/or type, as well as measurements of native glenoid version and inclination (ie, beta angle), to simulate the additional predictive accuracy that could be acquired through utilization of data readily available from computed tomography (CT)–based preoperative planning software.

Similarly to the methodology in our previous work,[12] these data were split 2:1 into mutually exclusive data sets to build and test the predictive models using each of the full and abbreviated feature sets for each outcome metric at each postoperative time point. A random selection of 66.7% of the data defined the training cohort, and the remaining 33.3% defined the validation test cohort. The performance of the full and abbreviated models to predict postoperative outcomes after aTSA and rTSA was quantified by the mean absolute error (MAE) between the actual and predicted values for each outcome measure at each postoperative time point in the 33.3% validation test cohort. To evaluate the relative learning ability of the full and abbreviated XGBoost models, we also conducted a baseline average analysis (ie, average error associated with selecting the cohort average) as the study control. Finally, a subgroup analysis was performed using the XGBoost machine learning technique for the full and abbreviated models to predict whether a patient would experience clinical improvement greater than the MCID[19] and SCB[20] patient satisfaction anchor-based thresholds for each outcome measure at 2-3 years of follow-up. The performance of the full and abbreviated models to predict whether a patient would achieve the MCID and SCB improvement thresholds was quantified using the classification metrics of precision (or positive predictive value, which quantifies the ability of a model to not identify a negative finding as a positive finding), recall (or sensitivity, which quantifies the ability of a model to identify a positive finding as a positive finding), and the area under the receiver operating characteristic curve (AUROC).[2,9,11] In the field of data science, an AUROC of 0.5 is considered random discrimination for a predictive model; >0.7 to 0.8, acceptable; >0.8 to 0.9, good; and >0.9, excellent.[9,11]

Results

The clinical data from 2153 primary aTSA patients (7305 visits; average follow-up period, 46.4 ± 35.6 months) and 3621 primary rTSA patients (10,122 visits; average follow-up period, 31.0 ± 25.8 months) were used to build and test predictive models at each postoperative time point: 3-6 months (1282 visits by aTSA patients and 2227 visits by rTSA patients), 6-9 months (658 and 1177 visits, respectively), 1 year (1451 and 2445 visits, respectively), 2-3 years (1347 and 1882 visits, respectively), 3-5 years (1321 and 1482 visits, respectively), and ≥5 years (1246 and 909 visits, respectively). A summary of demographic characteristics, diagnoses, and comorbidities for the aTSA and rTSA patient cohorts are presented in Table II. Preoperative outcomes, postoperative outcomes, preoperative-to-postoperative improvements in outcomes, and complication rates for the aTSA and rTSA patient cohorts at each follow-up point are presented in Tables III and IV, respectively. The aTSA and rTSA outcomes at each follow-up point stratified by age and sex are presented in Supplementary Tables S1-S16.

Table II Comparison of demographic characteristics, diagnoses, and comorbidities of primary aTSA and primary rTSA patients

	aTSA patients	rTSA patients
Demographic characteristic		
Age at surgery, yr	66.1 ± 9.2	72.5 ± 7.8
Sex: F/M/unknown	1111/1027/15	2350/1242/29
Height, in (cm)	66.6 ± 4.3 (169.2 ± 10.9 cm)	65.0 ± 4.0 (165.1 ± 10.2 cm)
Weight, lb (kg)	188.9 ± 44.5 (85.7 ± 20.2 kg)	172.7 ± 41.0 (78.3 ± 18.6 kg)
Body mass index	29.9 ± 6.3	28.7 ± 6.0
% with previous shoulder surgery	15.7	24.7
Surgery on dominant shoulder, %	55.8	62.2
Diagnosis, %		
Osteoarthritis	92.7	53.2
Osteonecrosis	3.2	2.4
Rotator cuff tear	2.6	39.3
Cuff tear arthropathy	0.8	38.4
Rheumatoid arthritis	3.1	3.5
Post-traumatic arthritis	2.1	2.4
Comorbidities, %		
No comorbidities	35.9	33.0
Inflammatory arthritis	11.7	7.8
Hypertension	47.4	53.3
Heart disease	13.6	16.2
Diabetes	12.2	13.7
Chronic renal failure	1.2	2.0
Tobacco use	9.8	7.2

aTSA, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty; *F*, female; *M*, male.

A comparison of the error between the actual and predicted outcomes in the validation data set demonstrates that both the full and abbreviated XGBoost models had lower MAEs relative to the baseline average study control MAEs for all clinical outcome measure predictions, as well as for both aTSA and rTSA at all postoperative time points (Table V). MAEs associated with each aTSA and rTSA outcome prediction at each postoperative time point are presented in Supplementary Tables S17-S23. As described in these tables, the prediction accuracy observed for aTSA and rTSA was similar for both the full and abbreviated models; however, for both the full and abbreviated models, MAEs were slightly higher at early postoperative time points than at later time points. A comparison of MAEs between the full and abbreviated models demonstrates that each model had similar predictive accuracy for each outcome measure, despite the abbreviated model only using a minimal feature set of preoperative inputs to inform its predictions: ASES score (±11.7 with full model vs. ±12.0

Table III Comparison of preoperative and postoperative outcomes and improvement in outcomes in primary aTSA patients analyzed in study, stratified by follow-up duration

	ASES score (postoperatively/ improvement)	Constant score (postoperatively/ improvement)	Global Shoulder Function score (postoperatively/ improvement)	VAS pain score (postoperatively/ improvement)	Active abduction (postoperatively/ improvement), °	Active forward elevation (postoperatively/ improvement), °	Active external rotation (postoperatively/ improvement), °	Adverse event rate, %/revision rate, %
Preoperative	35.7 ± 16.4	38.3 ± 14.3	4.1 ± 2.0	6.4 ± 2.1	83.1 ± 30.8	97.1 ± 32.3	19.9 ± 19.6	NA
Follow-up duration in aTSA patients								
3-6 mo	76.5 ± 18.9/ 40.4 ± 21.4	62.3 ± 15.5/ 24.4 ± 17.1	7.2 ± 2.2/ 3.0 ± 2.8	1.7 ± 2.1/ 4.7 ± 2.7	110.2 ± 32.0/ 26.3 ± 37.2	125.4 ± 33.5/ 29.3 ± 38.0	42.6 ± 18.7/ 22.4 ± 19.7	2.1/1.4
6-9 mo	82.4 ± 17.8/ 44.5 ± 21.5	68.1 ± 14.3/ 29.1 ± 15.9	7.7 ± 2.2/ 3.5 ± 2.7	1.3 ± 1.9/ 5.0 ± 2.7	119.8 ± 31.4/ 36.5 ± 37.2	136.0 ± 30.3/ 39.0 ± 36.2	49.5 ± 18.6/ 28.2 ± 20.8	1.5/0.8
1 yr	85.9 ± 17.1/ 49.2 ± 20.6	71.6 ± 13.6/ 33.1 ± 15.9	8.4 ± 1.9/ 4.3 ± 2.5	1.1 ± 1.9/ 5.3 ± 2.6	127.5 ± 32.0/ 44.8 ± 37.1	143.8 ± 29.0/ 47.3 ± 35.3	51.1 ± 18.7/ 30.7 ± 20.7	2.5/1.4
2-3 yr	87.4 ± 17.1/ 51.3 ± 20.9	73.4 ± 13.8/ 35.2 ± 16.0	8.6 ± 1.9/ 4.7 ± 2.5	1.0 ± 1.9/ 5.4 ± 2.6	129.1 ± 32.1/ 45.4 ± 39.1	146.5 ± 29.1/ 48.7 ± 37.2	52.4 ± 18.8/ 33.0 ± 22.0	3.4/2.0
3-5 yr	86.2 ± 17.4/ 50.5 ± 20.4	72.8 ± 13.6/ 35.5 ± 14.7	8.6 ± 2.0/ 4.5 ± 2.6	1.1 ± 1.9/ 5.3 ± 2.6	128.4 ± 31.6/ 44.3 ± 37.9	146.8 ± 29.2/ 49.4 ± 36.2	51.5 ± 18.9/ 31.8 ± 21.7	2.6/2.0
≥5 yr (average follow-up, 90.0 mo)	81.5 ± 20.3/ 47.1 ± 23.7	68.3 ± 14.7/ 32.7 ± 16.6	8.2 ± 2.2/ 4.4 ± 2.7	1.5 ± 2.3/ 5.0 ± 2.9	120.5 ± 31.9/ 37.3 ± 38.0	141.0 ± 31.6/ 44.8 ± 37.1	46.1 ± 20.0/ 31.1 ± 22.9	5.2/3.4
Full primary aTSA cohort								3.0/1.9

aTSA, anatomic total shoulder arthroplasty; *ASES*, American Shoulder and Elbow Surgeons; *VAS*, visual analog scale.

Table IV Comparison of preoperative and postoperative outcomes and improvement in outcomes in primary rTSA patients analyzed in study, stratified by follow-up duration

	ASES score (postoperatively/ improvement)	Constant score (postoperatively/ improvement)	Global Shoulder Function score (postoperatively/ improvement)	VAS pain score (postoperatively/ improvement)	Active abduction (postoperatively/ improvement), °	Active forward elevation (postoperatively/ improvement), °	Active external rotation (postoperatively/ improvement), °	Adverse event rate, %/revision rate, %
Preoperative	34.7 ± 15.9	34.9 ± 14.4	3.7 ± 2.1	6.3 ± 2.2	72.7 ± 36.9	85.7 ± 39.3	18.5 ± 21.3	NA
Follow-up duration in rTSA patients								
3-6 mo	73.4 ± 18.8/ 38.9 ± 20.5	58.3 ± 14.9/ 23.7 ± 16.8	6.9 ± 2.1/ 3.2 ± 2.8	1.7 ± 2.1/ 4.6 ± 2.7	104.9 ± 31.4/ 33.1 ± 39.6	120.9 ± 31.9/ 37.1 ± 42.1	28.5 ± 17.6/ 9.8 ± 21.4	3.1/0.7
6-9 mo	77.8 ± 18.1/ 41.2 ± 19.9	62.8 ± 13.7/ 28.2 ± 16.3	7.3 ± 2.0/ 3.7 ± 2.7	1.5 ± 2.0/ 4.5 ± 2.7	110.1 ± 29.7/ 39.4 ± 37.4	130.0 ± 29.1/ 46.6 ± 40.5	31.6 ± 17.7/ 13.6 ± 22.7	2.5/0.9
1 yr	81.2 ± 18.1/ 46.4 ± 20.7	67.0 ± 14.0/ 32.0 ± 16.2	7.9 ± 2.0/ 4.3 ± 2.6	1.3 ± 2.0/ 5.0 ± 2.7	120.6 ± 30.1/ 47.1 ± 39.7	137.8 ± 27.7/ 51.5 ± 41.2	35.7 ± 18.1/ 16.9 ± 22.2	2.3/1.0
2-3 yr	82.6 ± 18.1/ 46.7 ± 20.7	69.0 ± 13.8/ 34.1 ± 16.1	8.1 ± 1.9/ 4.4 ± 2.6	1.2 ± 2.0/ 5.0 ± 2.7	118.8 ± 30.6/ 46.5 ± 39.1	139.2 ± 26.9/ 54.2 ± 41.9	36.8 ± 17.6/ 18.8 ± 23.1	3.1/1.1
3-5 yr	82.2 ± 18.8/ 45.9 ± 21.6	68.0 ± 13.9/ 32.5 ± 15.7	8.2 ± 1.9/ 4.4 ± 2.7	1.2 ± 2.0/ 5.0 ± 2.7	117.9 ± 29.1/ 45.2 ± 39.1	137.3 ± 26.9/ 52.4 ± 41.7	36.3 ± 17.9/ 17.3 ± 23.4	2.7/1.3
≥5 yr (average follow-up, 80.6 mo)	79.9 ± 20.5/ 43.7 ± 23.5	65.7 ± 14.9/ 30.3 ± 17.3	7.9 ± 2.2/ 4.0 ± 2.9	1.4 ± 2.2/ 4.9 ± 2.9	112.3 ± 29.2/ 35.9 ± 39.6	130.7 ± 29.5/ 41.3 ± 42.9	32.3 ± 19.6/ 11.8 ± 25.3	2.5/1.1
Full primary rTSA cohort								2.7/1.0

rTSA, reverse total shoulder arthroplasty; *ASES*, American Shoulder and Elbow Surgeons; *VAS*, visual analog scale.

Table V Comparison of MAE associated with each outcome measure prediction for full and abbreviated machine learning models

Prediction error	Weighted average MAE (aTSA, rTSA)						
	ASES score	Constant score	Global Shoulder Function score	VAS pain score	Active abduction, °	Active forward elevation, °	Active external rotation, °
Baseline average	15.3 (15.6, 14.9)	12.5 (12.7, 12.5)	1.6 (1.6, 1.6)	1.6 (1.5, 1.6)	26.5° (26.4, 26.3)	23.0° (22.9, 23.2)	16.0° (16.2, 15.7)
XGBoost with full model	11.7 (11.6, 11.8)	8.9 (9.4, 9.1)	1.4 (1.4, 1.3)	1.3 (1.2, 1.3)	20.4° (20.9, 20.1)	17.6° (18.0, 17.8)	12.2° (13.1, 11.7)
XGBoost with abbreviated model	12.0 (11.9, 12.2)	9.8 (10.1, 9.9)	1.5 (1.5, 1.4)	1.4 (1.2, 1.5)	21.8° (22.0, 21.3)	19.2° (19.7, 19.2)	12.6° (13.2, 12.1)
MAE difference (full – abbreviated)	0.3 (0.3, 0.4)	0.9 (0.7, 0.8)	0.1 (0.1, 0.1)	0.1 (0.0, 0.2)	1.4° (1.1, 1.2)	1.6° (1.7, 1.4)	0.4° (0.1, 0.4)
XGBoost with abbreviated model plus implant data	12.0 (11.7, 12.0)	9.7 (9.8, 9.8)	1.4 (1.4, 1.4)	1.3 (1.2, 1.4)	21.7° (21.9, 21.1)	19.0° (19.2, 19.1)	12.4° (13.1, 12.0)

MAE, mean absolute error; *aTSA*, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty.
The presented MAE values are weighted averages over each postoperative time point (3-6 months, 6-9 months, 1 year, 2-3 years, 3-5 years, and ≥5 years); the MAE for each outcome measure at each postoperative time point is reported in Supplementary Tables S17-S23.

with abbreviated model), Constant score (±8.9 vs. ±9.8), Global Shoulder Function score (±1.4 vs. ±1.5), VAS pain score (±1.3 vs. ±1.4), active abduction (±20.4° vs. ±21.8°), forward elevation (±17.6° vs. ±19.2°), and external rotation (±12.2° vs. ±12.6°). Specifically, across all postoperative time points analyzed, the average difference in the MAE between the full and abbreviated model predictions was ±0.3 MAE for the ASES score (±0.3 in aTSA patients and ±0.4 in rTSA patients), ±0.9 for the Constant score (±0.7 and ±0.8, respectively), ±0.1 for the Global Shoulder Function score (±0.1 and ±0.1, respectively), ±0.1 for the VAS pain score (±0.0 and ±0.2, respectively), ±1.4° for abduction (±1.1° and ±1.2°, respectively), ±1.6° for forward elevation (±1.7° and ±1.4°, respectively), and ±0.4° for external rotation (±0.1 and ±0.4°, respectively). Of note, only marginal improvements in MAEs were observed for each outcome measure prediction when the abbreviated model was supplemented with data on implant size and/or type and measurements of native glenoid anatomy (Table V).

The full and abbreviated model predictions for MCID improvement in each outcome metric at 2-3 years of follow-up are presented in Table VI. The full predictive models achieved 82%-96% accuracy in the MCID with an AUROC between 0.75 and 0.97 for aTSA patients, whereas the abbreviated predictive models achieved 82%-96% accuracy in the MCID with an AUROC between 0.70 and 0.95 for aTSA patients. The full predictive models achieved 91%-99% accuracy in the MCID with an AUROC between 0.82 and 0.98 for rTSA patients, whereas the abbreviated predictive models achieved 91%-99% accuracy in the MCID with an AUROC between 0.84 and 0.94 for rTSA patients. Similarly, the full and abbreviated model predictions for SCB improvement in each outcome metric at 2-3 years of follow-up are presented in Table VII. The full predictive models achieved 79%-90% accuracy in SCB with an AUROC between 0.74 and 0.90 for aTSA patients, whereas the abbreviated predictive models achieved 76%-90% accuracy in SCB with an AUROC between 0.70 and 0.89 for aTSA patients. Finally, the full predictive models achieved 83%-92% accuracy in SCB with an AUROC between 0.78 and 0.88 for rTSA patients, whereas the abbreviated predictive models achieved 81%-90% accuracy in SCB with an AUROC between 0.70 and 0.87 for rTSA patients.

Discussion

The results of our 5774-patient machine learning study demonstrate that an abbreviated model using a minimal feature set of only 19 preoperative inputs provides similar accuracy to the full model using 291 inputs when predicting aTSA and rTSA outcomes at multiple postoperative time points. At each postoperative time point, the full and abbreviated models had similar MAEs when predicting each outcome measure, demonstrating the capability of the

Table VI XGBoost predictions using full and abbreviated models for aTSA and rTSA patients who experienced clinical improvement at 2 to 3 years' follow-up greater than MCID[19] threshold for multiple different outcome measures

MCID prediction	ASES score (aTSA, rTSA)	Constant score (aTSA, rTSA)	Global Shoulder Function score (aTSA, rTSA)	VAS pain score (aTSA, rTSA)	Abduction (aTSA, rTSA)	Forward elevation (aTSA, rTSA)	External rotation (aTSA, rTSA)
MCID	13.6 (17.0, 10.3)	5.7 (12.8, −0.3)	1.4 (1.7, 1.0)	1.6 (2.7, 1.4)	7.0° (13.9°, −1.9°)	12.0° (23.1°, −2.9°)	3.0° (14.5°, −5.3°)
Patient %	77.9 (72.9, 80.6)	71.3 (63.0, 77.6)	75.7 (75.5, 75.8)	75.3 (66.5, 77.9)	83.5 (78.2, 88.7)	79.9 (73.4, 92.7)	84.7 (77.8, 90.4)
Full model							
Precision, %	95 (94, 95)	96 (97, 98)	94 (96, 93)	92 (91, 91)	94 (91, 98)	92 (87, 99)	95 (90, 99)
Recall, %	99 (97, 99)	99 (99, 100)	96 (95, 98)	98 (97, 99)	94 (91, 98)	95 (88, 99)	96 (93, 99)
Accuracy, %	95 (94, 95)	97 (96, 99)	92 (93, 92)	93 (92, 91)	90 (86, 98)	89 (82, 99)	95 (92, 99)
AUC	0.90 (0.90, 0.88)	0.95 (0.97, 0.96)	0.88 (0.91, 0.87)	0.87 (0.89, 0.82)	0.83 (0.80, 0.98)	0.79 (0.75, 0.95)	0.83 (0.78, 0.95)
Abbreviated model							
Precision, %	91 (90, 91)	94 (93, 94)	93 (94, 91)	91 (89, 91)	90 (87, 95)	89 (90, 97)	89 (85, 94)
Recall, %	99 (97, 99)	99 (97, 99)	95 (94, 97)	96 (96, 97)	91 (89, 95)	95 (91, 98)	92 (88, 96)
Accuracy, %	93 (92, 93)	97 (96, 99)	90 (89, 91)	90 (88, 92)	84 (82, 94)	87 (87, 98)	87 (84, 97)
AUC	0.88 (0.87, 0.84)	0.94 (0.95, 0.93)	0.87 (0.89, 0.86)	0.87 (0.86, 0.84)	0.76 (0.72, 0.94)	0.72 (0.70, 0.89)	0.78 (0.73, 0.90)

aTSA, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty; *MCID*, minimal clinically important difference; *AUC*, area under curve.

Table VII XGBoost predictions using full and abbreviated models for aTSA and rTSA patients who experienced clinical improvement at 2 to 3 years' follow-up greater than SCB[20] threshold for multiple different outcome measures

SCB prediction	ASES score (aTSA, rTSA)	Constant score (aTSA, rTSA)	Global Shoulder Function score (aTSA, rTSA)	VAS pain score (aTSA, rTSA)	Abduction (aTSA, rTSA)	Forward elevation (aTSA, rTSA)	External rotation (aTSA, rTSA)
SCB	31.5 (37.6, 25.9)	19.1 (25.4, 13.6)	3.1 (3.9, 2.4)	3.2 (3.8, 2.6)	28.5° (36.1°, 19.6°)	35.4° (45.5°, 22.3°)	11.7° (20.1°, 3.6°)
Patient %	66.7 (57.3, 73.1)	62.3 (52.8, 71.7)	58.8 (57.4, 70.7)	62.8 (61.5, 72.5)	64.7 (55.2, 73.9)	61.3 (48.4, 75.7)	70.1 (64.6, 82.0)
Full model							
Precision, %	88 (89, 88)	91 (93, 94)	83 (91, 91)	85 (92, 88)	86 (80, 89)	85 (85, 91)	85 (79, 92)
Recall, %	93 (93, 94)	95 (88, 97)	93 (86, 90)	99 (92, 97)	86 (87, 87)	91 (88, 93)	89 (92, 93)
Accuracy, %	87 (89, 88)	91 (90, 92)	85 (87, 86)	86 (89, 88)	82 (81, 83)	84 (86, 88)	81 (79, 88)
AUC	0.84 (0.89, 0.81)	0.90 (0.90, 0.88)	0.84 (0.87, 0.84)	0.82 (0.86, 0.81)	0.81 (0.80, 0.78)	0.83 (0.87, 0.83)	0.76 (0.74, 0.78)
Abbreviated model							
Precision, %	87 (88, 86)	90 (91, 92)	82 (89, 91)	85 (92, 88)	82 (76, 84)	78 (74, 86)	82 (76, 89)
Recall, %	92 (92, 93)	95 (88, 97)	93 (86, 90)	99 (92, 97)	84 (83, 84)	88 (81, 91)	88 (89, 91)
Accuracy, %	87 (89, 88)	90 (88, 90)	84 (85, 86)	87 (90, 88)	80 (79, 81)	81 (82, 81)	79 (76, 84)
AUC	0.82 (0.86, 0.76)	0.89 (0.89, 0.87)	0.83 (0.86, 0.83)	0.85 (0.87, 0.82)	0.72 (0.74, 0.70)	0.74 (0.78, 0.70)	0.73 (0.70, 0.76)

aTSA, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty; *SCB*, substantial clinical benefit; *AUC*, area under curve.

predictive algorithms to account for outcome variability both during the recovery period and even into mid- and long-term follow-up as outcomes decline with age and deterioration. Additionally, no differences in accuracy were observed between aTSA and rTSA outcome predictions at any time point for either the full or abbreviated model, demonstrating that the prediction algorithms were equally effective for each prosthesis type. Only minor improvements in the abbreviated model predictions were observed after incorporating data on implant size and/or type and measurements of native glenoid version and inclination. Furthermore, the full and abbreviated models were equally effective at risk stratifying patients using only preoperative data, by accurately identifying patients at greatest risk of poor outcomes based on MCID thresholds (full model accuracy $>$ 82% with AUROC $>$ 0.75 vs. abbreviated model accuracy $>$ 82% with AUROC $>$ 0.70), as well as identifying patients most likely to achieve excellent outcomes based on SCB thresholds (full model accuracy $>$ 79% with AUROC $>$ 0.74 vs. abbreviated model accuracy $>$ 76% with AUROC $>$ 0.70) at 2-3 years of follow-up for all outcome measurements.

Preoperatively communicating the expected result from a proposed surgical treatment is an important component of informed consent. However, few surgeons can accurately predict the outcomes that a cohort of patients may achieve, nor do most know whether a particular patient will fare better or worse (and by how much) than the ''average'' patient. Machine learning–based predictive outcome algorithms are not a simple heuristic; rather, these computational techniques analyze large quantities of clinical data and consider numerous parameters to inform their evidence-based outcome predictions. As such, these predictions are clinically useful for shared decision making and can be used to more effectively communicate expected outcomes and better inform the risks and benefits of a surgical procedure. For the shoulder surgeon, an evidence-based tool that can accurately predict patient-specific outcomes after aTSA and rTSA from input of only 19 patient questions and active ROM measurements has many practical applications. First, these predictions can better align the surgeon's objectives for the procedure with those of the patient and establish more accurate expectations of what can be achieved with shoulder arthroplasty, given each patient's unique demographic characteristics, diagnoses, and comorbidities. Better alignment between the patient and surgeon on what can be achieved with shoulder arthroplasty may translate into improved patient satisfaction with the procedure.[7,8,15,17] Furthermore, these predictions can aid the shoulder surgeon in selecting the best technique or treatment for a particular patient based on a comparison of multiple different projected results, considering the patient's unique model inputs, as exemplified in this study by the comparative aTSA and rTSA outcome predictions. When these predictions are considered in a comparative manner, they function as a patient-specific tool to optimize the clinical outcomes of various techniques and treatment options. Additionally, consideration of these predictions relative to the age- and sex-stratified outcomes in aTSA and rTSA patients (Supplementary Tables S1-S16) can be used as a quality assessment metric to assess performance at each postoperative time point. Future work should perform additional predictive analyses to compare and quantify the impact of complications on model predictions, particular as it relates to false-positive predictions for the MCID and SCB, as well as create patient-specific predictive models for complication risk associated with each of the various techniques and treatment options.

More controversial is the use of a predictive tool to identify whether a specific patient is an appropriate candidate for an elective surgical treatment. The abbreviated model accurately identified $>$82% of patients who would achieve improvement greater than the MCID threshold and $>$76% of patients who would achieve improvement greater than the SCB threshold across all outcome metrics analyzed. Furthermore, the abbreviated model algorithms were associated with average MCID AUROC values of 0.82 for aTSA and 0.89 for rTSA and average SCB AUROC values of 0.85 for aTSA and 0.82 for rTSA. Thus, our AUROC results suggest that these predictive algorithms created from a minimal feature set have, on average, good ($>$0.8) to excellent ($>$0.9) discrimination of patients in the validation cohort to achieve MCID and SCB improvement. Although these predictions can be helpful to preoperatively identify patients who are good (or poor) candidates for these elective procedures, it must be acknowledged that each patient's needs are unique and different and that patient-specific requirements for pain relief and functional improvement may not align with the established MCID[19] or SCB[20] improvement thresholds. As such, this tool should not define which patient is eligible for surgical treatment; instead, it should be used to support treatment and never be misused to deny treatment.

Machine learning predictions using the full model—and its 291 feature inputs—have limited practical application for creating a decision-support tool that can be used in the typical clinical setting, given the substantial input burden on the patient, office staff, and health care provider. Fortunately, our results demonstrate that similar levels of predictive accuracy can be attained using as few as 19 patient questions and active ROM measurements. We observed minor improvements in predictive accuracy when the abbreviated model was supplemented with data on implant size and/or type and measurements of native glenoid version and inclination; these additional data have the potential to be seamlessly added to the model from CT-based preoperative planning software without additional input required by the office staff. Although future work is necessary to create and deploy the clinical software that utilizes these machine learning algorithms, the results of our study objectively demonstrate the efficacy of a minimal–feature set algorithm to predict aTSA and rTSA

outcomes at multiple postoperative time points. Furthermore, use of this minimal feature set composed of the most predictive inputs represents an opportunity for more efficient data collection and resource utilization, as it is inferred from our results that the majority of the preoperative data in the full model (including most of the questions from the 5 outcome metrics contained in the database: ASES, Constant, University of California–Los Angeles, Simple Shoulder Test, and Shoulder Pain and Disability Index scores) are superfluous, adding little additional predictive benefit to our model. Future work can apply these machine learning techniques to construct a new and more efficient shoulder arthroplasty–specific patient-reported outcome measure that eliminates inputs of little predictive value and only utilizes those patient questions found to be highly predictive of postoperative outcomes and/or patient satisfaction.

Aside from our previous work,[12] the use of machine learning to predict outcomes after shoulder arthroplasty is new, although a few studies have recently used machine learning to predict short-term complications after shoulder arthroplasty[10] and outcomes after hip[4] and knee[4,23] arthroplasty. Our machine learning analysis of aTSA and rTSA outcomes builds on previous work that used more traditional statistical techniques to compare aTSA and rTSA outcomes.[5,6,18-21] Our results demonstrated similar MAEs between aTSA and rTSA predictions for each clinical outcome measure at each postoperative time point; however, at earlier postoperative time points, we observed slightly higher MAEs than at later time points, despite having more data at those earlier time points. This finding is likely due to the greater variability in outcomes early owing to patients having different healing rates and perhaps also due to different methods and utilization of rehabilitation programs. As has been reported previously by Simovitch et al[21] and Levy et al,[13] aTSA and rTSA patients can continue to experience improvement for up to 2 years after surgery, after which improvement plateaus; these findings are consistent with our own observations for both the aTSA and rTSA cohorts in our data set (Tables III and IV). Machine learning algorithms improve and reduce error by learning with new data; hence, as additional clinical data are obtained, future work will refine these algorithms to further reduce model MAEs and improve predictive accuracy.

Our study has several limitations. First, 30 different sites and/or surgeons contribute to our clinical outcome database, and data from each site or surgeon inevitably contain some bias. As such, the derived models will also contain bias.[1,2,14,16] To reduce collection bias and input variability, all sites were trained to collect data using standardized data forms, and all completed forms were independently verified. Second, each of the surgeons who contributed clinical data are experienced shoulder specialists who have multiple years of experience with the prosthesis used in this study; as such, these predictions may not translate to less-experienced surgeons or to surgeons who have not completed the learning curve with this platform shoulder prosthesis. Third, our clinical database consists only of patients who elected to undergo shoulder arthroplasty, and these patients are primarily elderly, non-Hispanic white patients of European descent. For example, we do not collect data on individuals who are candidates but elect to forgo surgery because of comorbid illness or financial or personal reasons. Therefore, model predictions may not be representative of the outcomes achieved by patients of different demographic characteristics, regions, and ethnicities and/or races, and model predictions may be biased against patients too sick to safely undergo the procedure or patients whose condition was not sufficiently degenerative to undergo the procedure. Fourth, our models were developed from a data set of patients who underwent primary aTSA and primary rTSA with 1 platform shoulder prosthesis, in which patients with revisions, humeral fractures, endoprosthesis or hemiarthroplasty were excluded; therefore, model predictions may not be appropriate for patients with those excluded indications or other prosthesis types or designs. Fifth, our study used 1 tree-based machine learning technique to construct algorithms that quantify outcomes after shoulder arthroplasty; other techniques, such as deep learning, could achieve better predictive accuracy than XGBoost, as has been shown previously[12] using the wide and deep technique.[24] Despite slight improvements in predictive accuracy using the wide and deep technique, we used XGBoost in our study because its predictions are more interpretable, providing an F-score identification of the most meaningful parameters used by the model. Knowledge of the model input parameters driving up or down the patient-specific predictions can be clinically useful, particularly if those features are modifiable by the patient. Sixth, our clinical database, while extensive, contains some missing data; fortunately, XGBoost manages missing values and data sparsity well and imputes missing values on its own by minimizing the error rate for each tree as it learns. Finally, although we used a minimal feature set of 19 of the most predictive features in our database, there may be other features that are more predictive and clinically meaningful that were not included in the full model and are not currently collected in our clinical database. Future work should continue to refine the feature set to identify more clinically meaningful and highly predictive parameters that minimize the model MAE while also minimizing the user input burden, thereby ensuring that the decision-support tool can be efficiently implemented in the clinical setting.

Conclusion

Using a commercially available supervised machine learning technique to analyze a clinical database of 1 platform shoulder prosthesis, we constructed predictive algorithms using a full model (of 291 inputs) and an abbreviated model (of 19 inputs) and attained similar accuracy with each model to predict outcomes after shoulder arthroplasty at multiple postoperative time

points in our study of 2153 primary aTSA and 3621 primary rTSA patients. The abbreviated prediction model was supplemented with data on implant size and/or type and measurements of native glenoid version and inclination, which demonstrated that marginal improvements can be achieved when incorporating preoperative CT planning data. Finally, both the full and abbreviated model algorithms were able to preoperatively risk stratify patients based on improvement predictions greater than the MCID and SCB patient-satisfaction thresholds for each outcome measure analyzed in our study. These promising results demonstrate an efficient utilization of machine learning algorithms to predict clinical outcomes. Our findings using a minimal feature set of only 19 preoperative inputs suggest that this tool may be easily used during a surgical consultation to improve decision-making related to shoulder arthroplasty.

Disclaimer

Clinical data collection was sponsored by Exactech, Inc.

Vikas Kumar is employed by Ken Sci.

Christopher Roche is employed by Exactech.

Steven Overman is employed by Ken Sci.

Ryan Simovitch is a consultant for Exactech.

Pierre-Henri Flurin is a consultant for Exactech and receives royalties on products related to this article

Thomas Wright is a consultant for Exactech and receives royalties on products related to this article.

Joseph Zuckerman is a consultant for Exactech and receives royalties on products related to this article.

Howard Routman is a consultant for Exactech.

Ankur Teredesai is employed by Ken Sci.

Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jse.2020.07.042.

References

1. Ahmad MA, Eckert C, Teredesai A. Interpretable machine learning in healthcare. IEEE Intell Inform Bull 2018;1:1-6.
2. Cabitza F, Banfi G. Machine learning in laboratory medicine: waiting for the flood? Clin Chem Lab Med 2018;56:516-24. https://doi.org/10.1515/cclm-2017-0287
3. Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. Front Bioeng Biotechnol 2018;6:75. https://doi.org/10.3389/fbioe.2018.00075
4. 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;477:1267-79. https://doi.org/10.1097/CORR.0000000000000687
5. Friedman RJ, Cheung EV, Flurin PH, Wright T, Simovitch RW, Bolch C, et al. Are age and patient gender associated with different rates and magnitudes of clinical improvement after reverse shoulder arthroplasty? Clin Orthop Relat Res 2018;476:1264-73. https://doi.org/10.1007/s11999.0000000000000270
6. Friedman RJ, Eichinger J, Schoch B, Wright T, Zuckerman J, Flurin PH, et al. Preoperative parameters that predict postoperative patient reported outcome measures and range of motion with anatomic and reverse total shoulder arthroplasty. JSES Open Access 2019;3:266-72. https://doi.org/10.1016/j.jses.2019.09.010
7. Gonzalez Saenz de Tejada M, Escobar A, Bilbao A, Herrera-Espiñeira C, García-Perez L, Aizpuru F, et al. A prospective study of the association of patient expectations with changes in health-related quality of life outcomes, following total joint replacement. BMC Musculoskelet Disord 2014;15:248. https://doi.org/10.1186/1471-2474-15-248
8. Gonzalez Sáenz de Tejada M, Escobar A, Herrera C, García L, Aizpuru F, Sarasqueta C. Patient expectations and health-related quality of life outcomes following total joint replacement. Value Health 2010;13:447-54. https://doi.org/10.1111/j.1524-4733.2009.00685.x
9. Gortmaker SL, Hosmer DW, Lemeshow S. Applied logistic regression. Contemp Sociol 1994;23:159.
10. Gowd AK, Agarwalla A, Amin NH, Romeo AA, Nicholson GP, Verma NN, et al. Construct validation of machine learning in the prediction of short-term postoperative complications following total shoulder arthroplasty. J Shoulder Elbow Surg 2019;28:e410-21. https://doi.org/10.1016/j.jse.2019.05.017
11. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Assessing Fit of the Model. In: Hosmer DW Jr, Lemeshow S, Sturdivant RX, editors. Applied logistic regression. Hoboken, NJ: John Wiley & Sons; 2013. p. 177.
12. Kumar V, Roche C, Overman S, Simovitch R, Flurin PH, Wright T, et al. What is the accuracy of three different machine learning techniques to predict clinical outcomes after shoulder arthroplasty? Cline Orthop Relat Res 2020. in press. https://doi.org/10.1097/CORR.0000000000001263
13. Levy JC, Everding NG, Gil CC Jr, Stephens S, Giveans MR. Speed of recovery after shoulder arthroplasty: a comparison of reverse and anatomic total shoulder arthroplasty. J Shoulder Elbow Surg 2014;23:1872-81. https://doi.org/10.1016/j.jse.2014.04.014
14. Lipton ZC. The mythos of model interpretability. Queue 2018;16:31-57. https://queue.acm.org/detail.cfm?id=3241340. Accessed December 19, 2019.
15. Mahomed NN, Liang MH, Cook EF, Daltroy LH, Fortin PR, Fossel AH, et al. The importance of patient expectations in predicting functional outcomes after total joint arthroplasty. J Rheumatol 2002;29:1273-9.
16. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216-9. https://doi.org/10.1056/NEJMp1606181
17. Palazzo C, Jourdan C, Descamps S, Nizard R, Hamadouche M, Anract P, et al. Determinants of satisfaction 1 year after total hip arthroplasty: the role of expectations fulfilment. BMC Musculoskelet Disord 2014;15:53. https://doi.org/10.1186/1471-2474-15-53
18. Parsons M, Routman H, Roche C, Friedman R. Patient-reported outcomes of reverse total shoulder arthroplasty: a comparative risk factor analysis of improved versus unimproved cases. JSES Open Access 2019;3:174-8. https://doi.org/10.1016/j.jses.2019.07.004
19. Simovitch R, Flurin PH, Wright T, Zuckerman JD, Roche CP. Quantifying success after total shoulder arthroplasty: the minimal clinically important difference. J Shoulder Elbow Surg 2018;27:298-305. https://doi.org/10.1016/j.jse.2017.09.013
20. Simovitch R, Flurin PH, Wright T, Zuckerman JD, Roche CP. Quantifying success after total shoulder arthroplasty: the substantial clinical benefit. J Shoulder Elbow Surg 2018;27:903-11. https://doi.org/10.1016/j.jse.2017.12.014

21. Simovitch RW, Friedman RJ, Cheung EV, Flurin PH, Wright T, Zuckerman JD, et al. Rate of improvement in clinical outcomes with anatomic and reverse total shoulder arthroplasty. J Bone Joint Surg Am 2017;99:1801-11. https://doi.org/10.2106/JBJS.16.01387
22. Torlay L, Perrone-Bertolotti M, Thomas E, Baciu M. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain Inform 2017;4:159-69. https://doi.org/10.1007/s40708-017-0065-7
23. Twiggs JG, Wakelin EA, Fritsch BA, Liu DW, Solomon MI, Parker DA, et al. Clinical and statistical validation of a probabilistic prediction tool of total knee arthroplasty outcome. J Arthroplasty 2019;34:2624-31. https://doi.org/10.1016/j.arth.2019.06.007
24. Zheng Z, Yang Y, Niu X, Dai HN, Zhou Y. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. IEEE Trans Industr Inform 2017;14:1606-15. https://doi.org/10.1109/tii.2017.2785963