External Validation of a Machine Learning Based Clinical Outcomes Prediction Tool for aTSA and rTSA

Kumar V, Schoch B, Aibinder W, Parsons P, Watling J, Ko J, Gobbato B, Throckmorton T, Routman H, Polakovic S, and Roche C

Disclosures: V. Kumar: 3A; Exactech, Inc.; B. Schoch: 3B; Exactech, Inc. 7A; Exactech Inc.; Aibinder, W, 3B; Exactech, Inc. 7A; Exactech Inc.; Parsons, P, 3B; Exactech, Inc. 7A; Exactech Inc.; Watling, J, 3B; Exactech, Inc. 7A; Exactech Inc.; Ko, J, 3B; Exactech, Inc. 7A; Exactech Inc.; Gobbato, B, 3B; Exactech, Inc. 7A; Exactech Inc., Throckmorton, T, 3B; Exactech, Inc. 7A; Exactech Inc. Routman: 3B; Exactech, Inc. 7A; Exactech, Inc; S Polakovic: 3A; Exactech Inc. C. Roche: 3A; Exactech, Inc. 4; Exactech Inc.

Introduction: Predict+ (Exactech, Inc., Gainesville, FL) is a machine learning (ML) based clinical decision support tool (CDST) that uses a supervised ML algorithm (XGBoost) to preoperatively predict personalized clinical outcomes after anatomic (aTSA) and reverse (rTSA) total shoulder arthroplasty from a minimal feature set of 19 pre-operative inputs.[1-3] This CDST software provides personalized regression predictions for 7 outcome measures (VAS Pain, Global Shoulder Function, Shoulder Arthroplasty Smart Score, active abduction, active forward elevation, active external rotation, and internal rotation score) at 6 postoperative timepoints (3-6 months, 6-9 months, 1 year, 2-3 years, 3-5 years, and 5+ years) after both aTSA and rTSA. With additional inputs, the ASES and Constant score can also be predicted at the same timepoints. These supervised ML algorithms were developed from 2,270 primary aTSA and 4,198 primary rTSA patients from >30 different clinical sites; the internal validation of these algorithms has been published.[1-3] The aim of this study is to externally validate the Predict+ ML algorithms by comparing the accuracy of the predictions for patients who preoperatively received a personalized prediction report and compare their actual experienced results up to 2 years after surgery to what was originally predicted.

Methods: This external validation was performed on 243 patients (120F/123M) who received a Predict+® personalized report prior to surgery and had short-term clinical follow-up from 3 months to 2 years after primary aTSA (n=43) or rTSA (n=200). Specifically, the accuracy of the Predict+ predictions was quantified for the first four regression prediction timepoints (3-6 months, 6-9 months, 1 year, 2-3 years) for each aTSA and rTSA outcome model by comparing the mean absolute error (MAE) between each patient's preoperative prediction to each patient's actual clinical result. The MAE was calculated as the mean of every patient's absolute difference between the actual and predicted value for each outcome measure at each post-operative timepoint; MAE was calculated for the combined aTSA and rTSA cohort and also for each individual prosthesis type cohort. Finally, the accuracy of the external validation was compared to the accuracy published[1-3] from the internal validation for each outcome model.

Results: The predictive accuracy of the external validation is described by the MAE for each ML algorithm at each timepoint is presented in Table 1. Comparing the MAE associated with each 2 years outcome measure to the MAE from the internal validation[1-3] demonstrates that each ML models are generally performing as expected, with a few relatively small differences in accuracy between the internal and external validation results. Only a few of the ML models from the external validation performed worse than the internal validation, specifically the ASES and IR score predictions; however, every other ML model in the external validation was demonstrated to be more accurate than in the internal validation. A review of the distribution of MAE revealed some interesting findings. Notably, there was a tendency for the ML models to be more conservative with its predictions, as it was observed that patients averaged 5-10% better than predicted. Specifically regarding active range of motion predictions across all timepoints, 55.5% of patients achieved more abduction than predicted, 59.1% of patients achieved more forward elevation than predicted, 55.5% of patients achieved more external rotation than predicted, and 59.7% of patients achieved more internal rotation than predicted. Similarly regarding outcome score predictions across all timepoints, 51.3% of patients achieved more VAS pain improvement than predicted, 60.4% of patients achieved more shoulder function improvement than predicted, 53.8% of patients achieved more SAS score improvement than predicted, 53.4% of patients achieved more ASES score improvement than predicted, and 59.6% of patients achieved more Constant score improvement than predicted. Regarding the distribution of error for active range of motion predictions across all timepoints: 56.0% of abduction predictions were within 20° and 74.1% of abduction predictions were within 30°; 59.6% of forward elevation predictions were within 20° and 80.2% of forward elevation predictions were within 30°; 56.3% of external rotation predictions were within 10° and 75.2% of external rotation predictions were within 15°; and 54.2% of IR score predictions were within 1 point and 87.0% of IR score predictions were within 2 points. Similarly, regarding the distribution of error for outcome score predictions across all timepoints: 48.4% of VAS pain predictions were within 1 point and 78.4% of VAS pain predictions were within 2 points; 40.4% of global shoulder function predictions were within 1 point and 76.0% of global shoulder function predictions were within 2 points; 37.9% of SAS predictions were within 5 points, 71.0% of SAS predictions were within 10 points, and 86.2% of SAS predictions were within 15 points; 46.4% of ASES predictions were within 10 points, 66.0% of ASES predictions were within 15 points, and 78.3% of ASES predictions were within 20 points; 30.8% of Constant predictions were within 5 points, 56.8% of Constant predictions were within 10 points, and 76.0% of Constant predictions were within 15 points. Finally, there were 7 rTSA patients that serious complications (2 acromial/scapular fractures, 2 humeral fractures, 2 dislocations, and 1 case of unexplained pain); no aTSA patients had a complication. Interestingly, these 7 patients were responsible for some of the largest MAE outliers.

Discussion: Predict+ was released in November 2020 and has been used clinically to make personalized predictions on >3,500 patients since its launch in the US and select international markets. The results of this external validation of the first 4 post-operative prediction timepoints are promising and suggest that the predictive accuracy of this tool when used prospectively, is as good or better as that demonstrated in the internal validation. ML-based CDSTs have great potential to facilitate more evidence-based decision making and improve pre-operative patient counseling by helping patients better understand the actual risks and benefits associated with given treatment option. By better aligning patient expectations with realistic results, CDSTs can facilitate greater shared decision making between the patient and surgeon with the aim of achieving even better outcomes and greater levels of patient satisfaction. Before this potential can be full realized, it is necessary that the predictive accuracy of the CDST be demonstrated when prospectively used on the target population of patients. This external validation has numerous limitations. First, the sample size of only 243 patients is relatively small, particularly for aTSA, which only had 43 patients. Second, only the first 4 timepoints were evaluated because this software has not been clinically available long-enough for patients to have achieved a longer clinical follow-up duration. Third, this external validation only analyzed the regression predictions and did not evaluate the accuracy of the MCID and SCB classification predictions, which are predicted in this CDST at 2-3 years after surgery (when the patient has reached their full improvement/recovery). Future work should externally validate all post-operative timepoints, including each regression and classification model. Finally, future external validations should be performed on a patient cohort that is sufficiently large to evaluate the fairness and bias of the predictions relative to patients of protected sociodemographic attributes and also perform an outlier analysis.

Significance: This is the first orthopedic study to externally validate a machine learning-based clinical decision support tool for predicting outcomes after aTSA and rTSA. The results of this study demonstrate that this predictive tool is performing as intended with a few relatively small differences in accuracy between the internal and external validation results.

References:

1. Kumar V, Allen C, Overman S, Teredesai A, Simovitch R, Flurin PH, Wright TW, Zuckerman JD, Routman H, Roche C. Development of a predictive model for a machine learning–derived shoulder arthroplasty clinical outcome score. Seminars in Arthroplasty: JSES. 32 (2): 226-237, 2022.
2. Kumar V, Roche C, Overman S, Simovitch R, Flurin P-H, Wright T, et al. Using machine learning to predict clinical outcomes after shoulder arthroplasty with a minimal feature set. J Shoulder Elbow Surg. 2021 May;30(5):e225–e236.
3. Kumar V, Schoch BS, Allen C, Overman S, Teredesai A, Aibinder W, et al. Using machine learning to predict internal rotation after anatomic and reverse total shoulder arthroplasty. J Shoulder Elbow Surg. 2022 May;31(5):e234–e245.

Table 1. MAE from an External Validation of 243 patients (120F/123M) who received a Predict+ personalized report prior to surgery and had short-term clinical follow-up from 3 months to 2 years after primary aTSA (n=43) or rTSA (n=200). Results are compared relative to the MAE of the internal validation[1,2,3] of each predictive model.

Clinical Outcome Measures	Mean Absolute Error, 3-6 months	Mean Absolute Error, 6-9 months	Mean Absolute Error, 1 year	Mean Absolute Error, 2 years	Internal Validation MAE[1,2,3]	% Difference Between External & Internal Validation at 2 years
ASES - aTSA	11.1	6.5	10.7	1.1	11.9	90.8% better
ASES - rTSA	15.0	12.0	13.8	15.1	12.2	23.8% worse
ASES Combined	14.3	11.4	13.3	13.2	12.0	10.0% worse
Constant - aTSA	5.1	4.4	7.5	6.3	10.1	37.6% better
Constant - rTSA	8.8	6.1	8.7	6.8	9.9	31.3% better
Constant Combined	8.2	5.9	8.5	6.7	9.8	31.6% better
SAS - aTSA	8.3	6.9	5.8	4.2	8.2	48.8% better
SAS - rTSA	9.1	6.7	8.4	8.5	8.3	2.4% worse
SAS Combined	9.0	6.7	7.9	7.9	8.2	3.7% better
VAS Pain - aTSA	1.4	1.4	0.9	0.6	1.2	50.0% better
VAS Pain - rTSA	1.8	1.4	1.3	1.3	1.5	13.3% better
VAS Pain Combined	1.8	1.4	1.2	1.3	1.4	7.1% better
Global Shoulder Function - aTSA	1.2	0.9	1.4	1.3	1.5	13.3% better
Global Shoulder Function - rTSA	1.8	1.4	1.4	1.2	1.4	14.3% better
Global Shoulder Function Combined	1.7	1.4	1.4	1.2	1.5	20.0% better
Active Abduction – aTSA	19.1º	12.3º	35.2º	7.2º	22.0º	67.3% better
Active Abduction – rTSA	23.3º	15.3º	23.4º	18.3º	21.3º	14.1% better
Active Abduction Combined	22.5º	15.0º	25.5º	16.9º	21.8º	22.5% better
Active Forward Elevation – aTSA	17.7º	23.2º	17.1º	7.4º	19.7º	62.4% better
Active Forward Elevation – rTSA	21.8º	18.9º	19.7º	17.4º	19.2º	9.4% better
Active Forward Elevation Combined	21.1º	19.3º	19.3º	16.1º	19.2º	16.1% better
Active External Rotation – aTSA	12.5º	11.2º	9.8º	13.8º	13.2º	4.5% worse
Active External Rotation – rTSA	11.0º	8.4º	13.1º	9.7º	12.1º	19.8% better
Active External Rotation Combined	11.3º	8.7º	12.5º	10.2º	12.6º	19.0% better
IR Score – aTSA	1.21	0.58	0.75	0.35	1.1	68.2% better
IR Score – rTSA	1.14	1.02	1.19	1.28	1.2	6.7% worse
IR Score Combined	1.15	0.98	1.11	1.15	1.1	4.5% worse