Available online at www.sciencedirect.com
ScienceDirect
www.jsesarthroplasty.org

Use of machine learning to assess the predictive value of 3 commonly used clinical measures to quantify outcomes after total shoulder arthroplasty

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

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

ARTICLE INFO

Keywords:
Machine learning
Clinical outcome measures
aTSA and rTSA outcomes
Shoulder arthroplasty
Predictive value

ABSTRACT

Background: An important psychometric parameter of validity that is rarely assessed is predictive value. In this study we utilize machine learning to analyze the predictive value of 3 commonly used clinical measures to assess 2-year outcomes after total shoulder arthroplasty (TSA).
Methods: XGBoost was used to analyze data from 2790 TSA patients and create predictive algorithms for the American Shoulder and Elbow Surgeons (ASES), Constant, and the University of California Los Angeles (UCLA) scores and also quantify the most meaningful predictive features utilized by these measures and for all questions comprising each measure to rank and compare their value to predict 2-year outcomes after TSA.
Results: Our results demonstrate that the ASES, Constant, and UCLA measures rarely considered the most-predictive features relevant to 2-year TSA outcomes and that each outcome measure was composed of questions with different distributions of predictive value. Specifically, the questions composing the UCLA score were of greater predictive value than the Constant questions, and the questions composing the Constant score were of greater predictive value than the ASES questions. We also found the preoperative Shoulder Pain and Disability Index (SPADI) score to be of greater predictive value than the preoperative ASES, Constant, and UCLA scores. Finally, we identified the types of preoperative input questions that were most-predictive (subjective self-assessments of pain and objective

Institutional review board approval was obtained per the institutional review board-approved protocol (NYU 57 IRB study # i05-144_MOD41).

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

https://doi.org/10.1053/j.sart.2020.12.003
1045-4527/

measurements of active range of motion and strength) and also those that were least-predictive of 2-year TSA outcomes (subjective task-specific activities of daily living questions).
Discussion: Machine learning can quantify the predictive value of the ASES, Constant, and UCLA scores after TSA. Future work should utilize this and related techniques to construct a more efficient and effective clinical outcome measure that incorporates subjective and objective input questions to better account for the preoperative factors that influence postoperative outcomes after TSA.
Level of Evidence: Level III; Retrospective Comparative Study

Clinical outcome measures quantify preoperative patient status and improvement after treatment using multiple subjective and objective assessments. Numerous outcome measures are used clinically in the shoulder, though no gold standard tool currently exists. The American Shoulder and Elbow Surgeons (ASES) score is among the most common assessment tools utilized in the United States to quantify clinical outcomes after shoulder surgery. The ASES score is a 0-100 point scoring system (100 = best score) developed in 1993 by the ASES research committee as a baseline measure of shoulder function that is applicable to all patients regardless of diagnosis [27]. The original ASES score consisted of 11 subjective patient survey questions, composed of 1 visual analog score pain assessment accounting for 50% and 10 activities of daily living (ADL) questions accounting for the remaining 50% [27]. The ASES was later modified (m-ASES) to remove 2 questions and add 4 new questions related to the hand/wrist to adapt the scoring system for the entire upper extremity [6]. Similarly, the Constant score is among the most common shoulder clinical outcome measures utilized in Europe. The Constant score is a 0-100 point scoring system (100 = best score) published in 1986 and is composed of 65% physical assessment (25% strength + 40% range of motion [ROM]) and 35% subjective patient assessment [10,11]. The University of California Los Angeles (UCLA) score is one of the oldest scores utilized to quantify clinical outcomes in the shoulder. It was published in 1981 and is a 0-35 point scoring system (35 = best score) that measures five different domains including pain (10 points), function (10 points), forward flexion (5 points), forward flexion strength (5 points), and patient satisfaction (5 points) [1]. Each clinical outcome measure is composed of different, but similar questions and each allocates a different scoring weight for shoulder pain, function, ROM, and strength. Despite differences, previous clinical research has demonstrated that these 3 outcome measures are strongly correlated ($R > 0.8$) when quantifying clinical outcomes after anatomic total shoulder arthroplasty (aTSA) and reverse total shoulder arthroplasty (rTSA) [14,24].

The psychometric properties of the ASES [3-7,14,16,20,21,24-27,30,31,34], Constant [9-12,14,19,20,31,34], and UCLA [14,20,34] outcome measures has been previously evaluated for different shoulder pathologies. While those analyses of reliability, validity, and responsiveness are necessary and essential, recent advances in clinical research and data science present a new characteristic of validity by which to evaluate these clinical instruments: predictive value. An emerging application of machine learning is to quantify and rank the predictive value of an outcome measure and each of its input questions based upon its utility to an algorithm trained to predict that measure. Specifically, by comparing the predictive value of an outcome measure and each of its input questions to the most meaningful features driving a predictive model of that measure, the relative importance of each input question can be assessed. Doing so can yield helpful information about what type of preoperative data most influence outcomes after total shoulder arthroplasty (TSA).

Recent work has utilized machine learning to create predictive algorithms for the ASES, Constant, and UCLA scores at various postoperative timepoints after aTSA and rTSA [22,23]. A detailed investigation of the most meaningful predictive features utilized by these machine learning models for each outcome measure and all questions comprising each measure will permit an objective assessment of the predictive value of the ASES, Constant, and UCLA scores. Therefore, the goal of this study is to utilize machine learning to quantify and compare the predictive value of the ASES, Constant, and UCLA outcome measures after TSA.

Methods

We utilized the XGBoost [33] machine learning technique to analyze a multicenter clinical outcomes database of shoulder arthroplasty patients and create predictive algorithms for the ASES, Constant, and UCLA scores at 2-years follow-up after TSA. This database consists of clinical outcomes collected prospectively from 30 different sites utilizing 1 platform shoulder prosthesis (Equinoxe, Exactech, Inc, Gainesville, FL, USA). All data was collected using standardized forms at each of the 30 clinical sites according to an institutional review board-approved protocol. All primary aTSA and primary rTSA patients in the database that were performed between November 2004 and April 2018 and had 18-36 months follow-up were included in this analysis to create the 2-year predictive models of the ASES, Constant, and UCLA scores. To ensure a homogenous dataset, patients with revisions, humeral fractures, endoprostheses, hemiarthroplasty, and also patients with postoperative visits <18 and >36 months follow-up were excluded. It should be noted that our study analysis focused only on 2 year outcomes in order to limit confounding variables as patients should have achieved full clinical improvement by this time while not yet experience the deteriorating effects that may be associated with longer-term follow-up.

XGBoost is a supervised, ensemble machine learning technique of multiple-regression trees that are built by iteratively partitioning the training dataset into multiple small batches using a method called boosting [33]. The predictive model utilized 291 inputs from the database, including demographics, diagnoses, comorbidities, implant type, ROM, radiographic findings, clinical outcome scores and the individual questions used to derive 5 different outcome measures, including ASES, Constant, UCLA, Simple Shoulder Test, and the Shoulder Pain and Disability Index (SPADI). Similar to our previous work [22,23], predictive models were created by splitting the database 2:1 into mutually exclusive datasets to build and test the ASES, Constant, and UCLA algorithms. A random selection of 66.7% of the data defined the training cohort and the remaining 33.3% defined the validation test cohort.

The ASES, Constant, and UCLA predictive algorithms were analyzed to identify and rank the preoperative input model features based upon their predictive value to each 2-year machine learning model. Specifically, all 291 features utilized in the database were ranked for each of the ASES, Constant, and UCLA algorithms according to their predictive value using the F-Score [33] and the Reciprocal Fusion Rank Score [13]. The F-Score is determined by the XGBoost machine learning technique and quantifies the predictive value of an individual feature to the overall algorithm by the frequency that each feature is used as a candidate for the split by the decision-tree algorithm [33]. The Reciprocal Fusion Rank Score combines the F-Score predictive value with the prevalence and uniqueness of that feature in the dataset, deprioritizing features with nonunique and also sparse inputs [13]. Of note, the feature uniqueness is computed using an information theory metric known as entropy, which measures the overall randomness and uncertainty of a feature's value across patients in the dataset [32]. The F-Score and Reciprocal Fusion Rank Score distribution associated with the preoperative ASES, Constant, and UCLA questions were quantified and compared using a 1-sided Wilcoxon Rank Sum Test. The null hypothesis for Wilcoxon Rank Sum Test is that distributions of F-Scores or Reciprocal Fusion Rank Scores between each outcome measures are equal. The alternative hypothesis is that an outcome measure distribution is either positively or negatively shifted relative to the distributions of the other measures. The significance level was 0.05. Finally, the F-Score and Reciprocal Fusion Rank Scores associated with each outcome measure were compared to the top 20 feature inputs used by each of the ASES, Constant, and UCLA predictive algorithms as an objective assessment of the predictive value of the features within each 2-year predictive model.

Results

Preoperative, intra-operative, and postoperative data from 2,790 patients (1141 aTSA, 1,649 rTSA) with 3,229 postoperative follow-up visits (1,347 aTSA, 1882 rTSA) were used to create predictive algorithms for the ASES, Constant, and UCLA scores. The F-Score and Reciprocal Fusion Rank Score associated with the questions composing the ASES (Table 1), Constant (Table 2), and UCLA (Table 3) outcome measures are presented in Tables 1, 2, and 3, respectively. Comparing the distribution of F-Scores and Reciprocal Fusion Rank Scores (Table 4) between the 3 outcome measures demonstrates the Constant F-scores are positively shifted relative to the distribution of ASES F-Scores (P= .0004) and the UCLA Reciprocal Fusion Rank Scores are positively shifted relative to the distribution of Constant Reciprocal Fusion Rank Scores (P= .0370; Table 4). A review of the F-Scores demonstrates most input questions composing the ASES and Constant scores were of low predictive value to each 2-year predictive model. Generally, the subjective self-assessments of pain and objective measurements of active ROM and strength were the preoperative questions of the greatest predictive value and conversely, the ADL input questions related to a patient's capability to perform a specific task were the preoperative questions of the lowest predictive value to 2-year TSA outcomes.

Table 1 – F-score and reciprocal rank score analysis of the individual ases score questions in the 2 year prediction model.

ASES score questions	F-score	RECIPROCAL fusion rank score
Preop pain on a daily basis	1811	0.031
Preop comfort of sleep on affected side	214	0.028
Preop reach a high shelf	206	0.028
Preop do usual activities/work	202	0.029
Preop put on a coat	199	0.028
Preop comb hair	195	0.029
Preop personal hygiene and toilet needs	171	0.028
Preop do usual recreational sport	171	0.028
Preop wash back/fasten bra	161	0.027
Preop throw ball overhand	113	0.026
Preop lift 10 lbs above shoulder	84	0.026
Mean ± standard deviation	321 ± 496	0.028 ± 0.001

A comparative analysis of the top 20 most meaningful feature inputs for each 2-year model demonstrates that only 1 of the top 20 predictive inputs to the ASES algorithm were associated with ASES score (Table 5), only 3 of the top 20 predictive inputs to the Constant algorithm were associated with the Constant score (Table 6), and only 4 of the top 20 predictive inputs to the UCLA algorithm were associated with the UCLA score (Table 7). Interestingly, the preoperative ASES, Constant, and UCLA scores were all observed to be of high predictive value even though, especially for ASES and Constant, its constituent questions were observed to be of low predictive value. However, the clinical outcome measure observed to be of the greatest predictive value was the preoperative SPADI score, as demonstrated by the high F-Scores and high Reciprocal Fusion Rank Scores in each of Tables 5-7, where each of these values were greater (and hence more predictive) to each of the 2-year ASES, Constant, and UCLA algorithms than the preoperative value of each score, respectively.

Discussion

The results of this study demonstrate that machine learning can be used to quantify the predictive value of the ASES,

Table 2 – F-score and reciprocal rank score analysis of the individual constant score questions in the 2 year prediction model.

Constant score questions	F-score	Reciprocal fusion Rank Score
Preop active abduction	4733	0.039
Preop active forward elevation	3646	0.038
Preop pain daily basis	1739	0.032
Preop max weight/strength assessment	1710	0.030
Preop move arm to top of head?	954	0.029
Preop move dorsum hand to lumbrosacral junction?	782	0.029
Preop move arm to waist?	779	0.023
Preop move arm above head?	672	0.028
Preop move arm to neck?	594	0.025
Preop move dorsum of hand to buttocks?	554	0.025
Preop move dorsum of hand to waist? (3rd lumbar vertebra)	530	0.026
Preop move arm/hand behind head with elbow held back?	528	0.027
Preop move arm/hand behind head with elbow held forward?	516	0.027
Preop move arm/hand to top of head with elbow held forward?	352	0.026
Preop move arm to xiphoid	325	0.022
Preop move dorsum of hand to 12th dorsal vertebra	279	0.024
Preop move arm full elevation	260	0.022
Preop move arm/hand top head with elbow held back	243	0.025
Preop comfort of sleep/unaffected sleep?	233	0.029
Preop do usual activities/work	231	0.029
Preop move dorsum of hand to interscapular region	204	0.022
Preop more dorsum of hand to lateral thigh	172	0.021
Preop do full recreational sport	170	0.028
Mean ± standard deviation	879 ± 1140	0.027 ± 0.005

Constant, and UCLA scores as well as the predictive value of the individual questions that compose each measure. We found, based on distribution differences in the F-Score and/or Reciprocal Fusion Rank Scores between measures, that the input questions composing the UCLA outcome measure are, as a whole, of greater predictive value than that of the Constant score and the input questions composing the Constant outcome measure are, as a whole, of greater predictive value than that of the ASES score pertaining to 2-year outcomes after TSA. Our study also demonstrated the majority of ASES and Constant questions were of low predictive value to the 2-year TSA predictive models. Despite this, the aggregate preoperative outcome scores were of high predictive value, with each score utilized in the top 20 most-predictive feature inputs for at least 2 of the 3 models. Furthermore, our analysis demonstrated that the objective measures of ROM and strength, and the subjective assessments of pain were among the most-predictive types of input questions, whereas the task-specific ADL input questions were of the lowest predictive value to each 2-year model.

Table 3 – F-score and reciprocal rank score analysis of the individual ucla score questions in the 2 year prediction model.

UCLA score questions	F-score	Reciprocal fusion Rank Score
Preop active forward flexion	2664	0.037
Preop function score	2043	0.033
Preop pain assessment	1444	0.032
Preop strength of forward flexion	101	0.026
Satisfaction (NA for preoperative assessment)	NA	NA
Mean ± standard deviation	1563 ± 1095	0.032 ± 0.005

A detailed review of the top 20 most-predictive features driving the ASES, Constant, and UCLA 2-year TSA models demonstrates the most-predictive inputs were rarely considered by any of the ASES, Constant, or UCLA outcome measures. Additionally, the shoulder outcome measure found to be of the greatest predictive value to each 2-year ASES, Constant, and UCLA algorithm was the preoperative SPADI score. The importance of the aggregate preoperative outcome score to the 2-year postoperative result aligns well with the recent findings of Friedman et al who used a multiple linear regression model with backward stepwise selection to identify the preoperative factors that influence postoperative outcomes for multiple different outcome measures after TSA [15]. Similar to our findings, they reported that the preoperative ASES score significantly influenced the postoperative ASES score for both aTSA and rTSA [15]. However, Friedman et al did not analyze the influence of the preoperative SPADI score on postoperative outcomes, nor did they assess the influence of the individual questions composing each outcome measure. Future work should quantify which characteristics of the SPADI account for its superior predictive performance with TSA.

Our findings suggest an opportunity for improvement in both efficiency and effectiveness with ASES, Constant, and UCLA outcome measures when quantifying TSA outcomes, and may also suggest the need for an altogether new clinical assessment tool that better accounts for the preoperative factors that influence postoperative outcomes after TSA. The existence of >25 different shoulder clinical outcome measures [4,18] and the current lack of consensus of a gold-standard measure further suggests the need for a new clinical outcome measure [2], particularly for TSA outcomes given the high cost of treatment and unique characteristics of the patient population. More efficient and effective clinical outcome measures are increasingly necessary given the quality assessment requirements associated with new value-based models and bundled payment initiatives, as quantifying clinical improvement is a critical component of the cost/benefit equation. Furthermore, the future will demand and even greater focus on outcome quality using patient-centered tools, as healthcare treatment decision-making becomes increasingly more shared.

The UCLA, Constant, and ASES outcome measures were developed in 1981, 1986, and 1993, respectively; our knowledge of clinical research, data science, and shoulder pathologies and treatment modalities have expanded significantly since these tools were deployed. It is critical that we continue to improve our tools, and if these historical outcome measures are not made more efficient and effective, then

Table 4 – Comparison of mean F-score and reciprocal rank score distributions for the outcome measure questions used in the ASES, constant, and UCLA prediction models, where $P<.05$ denotes a significant difference.

(Mean ± std dev)	ASES	Constant	UCLA
F-score	321 ± 496	879 ± 1140	1563 ± 1095
Reciprocal Fusion rank score	0.028 ± 0.001	0.027 ± 0.005	0.032 ± 0.005
P value (ASES vs. constant)	P= **.0004** (F-Score), P = .141 (Reciprocal Rank Score)		
P value (ASES vs. UCLA)	P = .085 (F-Score), P = .062 (Reciprocal Rank Score)		
P value (constant vs. UCLA)	P = .168 (F-Score), P = **.037** (Reciprocal Rank Score)		

Table 5 – Top 20 most predictive preoperative features used in the ASES 2 year prediction model.

Top 20 Most Meaningful Pre-operative Parameters for the 2yr ASES Score Prediction	F-Score	Reciprocal Fusion Rank Score	Included in ASES Score?
Follow-up duration	17826	0.044	No
Preop SPADI score	5173	0.038	No
Surgery on Dominant Hand?	4957	0.039	No
Preop active abduction	4919	0.038	No
Preop composite rom score	4817	0.039	No
Preop active external rotation	4063	0.037	No
Preop ASES score	3972	0.037	Yes
Preop active forward elevation	3812	0.037	No
Preop constant score	3656	0.036	No
Is gender female?	3546	0.038	No
Preop passive external rotation	3381	0.036	No
Preop internal rotation (IR) score	2972	0.034	No
Preop UCLA score	2880	0.034	No
Preop external rotation lag	2812	0.033	No
Preop SST score	2653	0.033	No
Preop pain when lying on affected side	2585	0.033	No
Comorbidity of hypertension	2558	0.033	No
Preop shoulder function	2558	0.032	No
Diagnosis of osteoarthritis	2546	0.033	No
Preop pain touching back of neck	2453	0.032	No

Table 6 – Top 20 most predictive preoperative features used in the constant 2 year prediction model.

Top 20 most meaningful pre-operative parameters for the 2 yr constant score prediction	F-score	Reciprocal fusion rank score	Included in Constant score?
Follow-up duration	16958	0.044	No
Preop SPADI score	4944	0.038	No
Preop active abduction	4733	0.039	Yes
Preop composite rom score	4684	0.040	No
Surgery on dominant hand?	3810	0.039	No
Preop active external rotation	3734	0.038	No
Preop constant score	3670	0.037	Yes
Preop active forward elevation	3646	0.038	Yes
Preop ASES score	3497	0.036	No
Preop passive external rotation	3242	0.036	No
Preop UCLA score	2933	0.034	No
Preop pain with lying on affected side	2898	0.033	No
Preop External Rotation Lag	2777	0.034	No
Preop pain when touching back of neck	2652	0.033	No
Preop global shoulder function score	2649	0.033	No
Preop IR score	2624	0.033	No
Preop SST score	2431	0.033	No
Comorbidity of hypertension	2414	0.033	No
Preop pain when pushing with affected arm	2172	0.032	No
Previous surgery?	2108	0.031	No

new and better measures [2] should be developed. Our study demonstrates that the constituent questions of the ASES, Constant, and UCLA scores are of low predictive value to 2-year TSA outcomes, and it should also be recognized that these measures have documented psychometric issues, such as the >20% postoperative ceiling effects with ASES score [20,31], the poor reliability and lack of standardization of the strength assessment with the Constant score

Table 7 – Top 20 most predictive pre-operative features used in the UCLA 2 year prediction model.

Top 20 most meaningful preoperative parameters for the 2 yr UCLA score prediction	F-score	Reciprocal fusion rank score	Included in UCLA score?
Follow-up duration	10214	0.045	No
Preop composite rom score	4194	0.041	No
Preop SPADI score	3982	0.038	No
Preop active abduction	3166	0.038	No
Preop constant score	2944	0.037	No
Preop ASES score	2784	0.037	No
Preop passive external rotation	2704	0.037	No
Preop active forward elevation	2664	0.037	Yes
Preop active external rotation	2576	0.036	No
Preop UCLA score	2346	0.035	Yes
Preop pain with lying on affected side	2074	0.034	No
Preop shoulder function	2043	0.033	Yes
Surgery on dominant hand?	1985	0.037	No
Preop external rotation lag	1906	0.033	No
Preop internal rotation (IR) score	1840	0.033	No
Preop pain when touching back of neck	1838	0.033	No
Preop SST score	1770	0.033	No
Preop pain when pushing with affected arm	1747	0.033	No
Is gender female?	1648	0.036	No
Preop pain on a daily basis	1444	0.032	Yes

[9,19,29], and also the age and gender bias with the Constant score when used for the typical TSA patient, as demonstrated by the multiple different age and gender normalization techniques [10,12,35,36].

New clinical assessment tools can be made more effective by selecting only the most valid questions that both reflect the patient perception of their health and treatment while also accounting for the preoperative factors that influence postoperative outcomes. The novel machine learning technique presented in this study is perhaps the best method to objectively quantify the predictive value of different input questions utilized in different outcome measures. Kumar et al [23] demonstrated how machine learning techniques can facilitate identification of a “minimal feature set” consisting only of the most meaningful predictive features. This minimal feature set identified a combination of both subjective and objective preoperative inputs to construct aTSA and rTSA postoperative predictive models for the visual analog score Pain, Global Shoulder Function, and 3 difference measures of active ROM [23]. Future work should attempt to adapt this minimal feature set of inputs to construct a more efficient and effective TSA-specific clinical outcome measure [2].

New clinical assessment tools can be made more efficient by reducing the overall number of questions as doing so will reduce administrative burden and responder fatigue while at the same time improve patient compliance. Minimizing the number of subjective questions asked to the patient is a goal of recent NIH-funded efforts to develop the Patient Reported Outcomes Measurement Information System (PROMIS). PROMIS is a patient-centered computer adaptive test that quantifies multiple domains of health measures for different target populations. The PROMIS Upper Extremity (UE) consists of a 46-question item bank that is a subset of the overall physical health assessment; a short form consisting of 7 static questions is also available. The computer adaptive testing algorithm dynamically responds to individual patient answers by filtering out nonrelevant questions to improve precision, so that different questions are administered to different patients even though all patients receive a score on the same scale. Typically, a patient may only be required to answer 3-5 questions [8,37]and Minoughan et al demonstrated that PROMIS UE required only 61 seconds to complete, which was significantly faster than the Simple Shoulder Test (93 seconds) and ASES measures (142 seconds) [26]. PROMIS potentially represents a significant advance in clinical research, but ultimately its efficacy will be determined by the validity of the questions in its item bank. The relevance of our study is demonstrated by a review of the PROMIS UE Item Bank 2.0, which consists almost exclusively of task-specific ADL questions. These task-specific questions closely resemble the ADL questions utilized by the ASES and Constant score that we identified as being of the lowest predictive value to 2-year TSA outcomes. While future work is necessary to quantify the predictive value and validity of the questions in the PROMIS item banks, based on our findings, it is unlikely that questions regarding “passing a 20 lb ham around a table” (a question used in the dynamic test and short form) will be predictive of outcome success after TSA. Though, a computer adaptive algorithm in combination with a machine learning optimized item bank constructed of the most-predictive questions, like the aforementioned “minimal feature set” for TSA [23], may represent the next great innovation in clinical research.

Our study has several limitations. First, we utilized only 1 machine learning technique (XGBoost) to quantify the predictive validity of each individual outcome measure question; other machine learning techniques, such as Random Forest as previously performed by Gowd et al [17] and Roche et al [28], may identify different most-predictive features. Second, we only analyzed the ASES, Constant, and UCLA outcome measures and questions based on their ability to predict short-term TSA outcomes, as defined as 18-36 months;

different most-predictive input features may be identified for different postoperative timepoints, such as the 3-6 month model or 5 year+ models previously developed by Kumar et al [22,23]. Third, while the 291 parameters utilized in our predictive models are numerous, our dataset is not exhaustive of all the possible parameters and it is very likely that there are additional features that are more predictive and more clinically meaningful, which are not currently collected in our database. For example, Gowd et al used machine learning to predict short-term complications and reported different most-predictive features, including: patient BMI, preoperative hematocrit, operating time, patient age, and preoperative albumin [17]. We did not observe that patient BMI, operating time, or patient age were in our top 20 most predictive features for any of the ASES, Constant, or UCLA algorithms, and our database did not contain hematocrit or albumin measures. Future work should continue to expand the scope of our clinical data collection efforts to include new parameters that may infer additional predictive value. Fourth, our F-Score and Reciprocal Fusion Rank Score analyses did not directly incorporate the different scoring weight allocations utilized by the ASES, Constant, and UCLA score calculations. For example, the ASES calculation prioritizes the subjective pain assessment as 50% of the overall score and allocates the remaining 50% to the 10 ADL questions [27]. Similarly, the Constant score allocates 15% to the 1 pain score, 20% to the 8 ADL questions, 20% for 2 goniometer ROM measurements, 20% for 11 different functional arm/hand positioning questions, and 25% for 1 power/strength question [10,11]. Thus, not all questions contribute equally to the aggregate score and our F-Score and Reciprocal Fusion Rank Score analysis assumed each question was of equal value. However, it was observed that the ASES and Constant questions that carried greater scoring weights also had greater F-Scores and Reciprocal Fusion Rank Scores and this finding likely accounts for why the aggregate preoperative scores were of greater predictive value than the average F-Scores of its individual input questions. Fifth, we did not assess the m-ASES [6] in this study as we are only studying the shoulder (and not the distal upper extremity); however, it is interesting to note that the 2 questions that were removed from the original ASES: (1) sleep on your painful side and (2) throw a ball overhand, were identified by our ASES predictive model as the second best overall and the second worst overall patient questions according to their F-Score values, respectively. And finally sixth, in this study we did not separately evaluate aTSA and rTSA, as previous work demonstrated that aTSA and rTSA predictive models had similar predictive accuracy for this patient population [22,23]. It may be that the most-predictive features driving the aTSA model and the rTSA model are different, and if so, these most meaningful features could be of a different rank-order and could also consist of different features altogether. Furthermore as these outcome measures are equally useful for both aTSA and rTSA applications, and also for other shoulder treatment options, it is appropriate for this initial analysis to combined the aTSA and rTSA cohort in order to assess the predictive validity of each outcome measure. Future work should identify and compare the most meaningful features driving aTSA and rTSA models for each of the ASES, Constant, UCLA, and SPADI outcome measures.

Conclusion

This machine learning analysis of the ASES, Constant, and UCLA clinical outcome metrics, using data from 2790 TSA patients, quantified the predictive value of each question from each measure based on its ability to predict 2-year TSA outcomes. Using this novel technique, we demonstrated that the UCLA questions were of greater predictive value than the Constant questions, and the Constant questions were of greater predictive value than the ASES questions. Additionally, we identified the types of preoperative input questions that were most-predictive (subjective self-assessments of pain and objective measurements of active ROM and strength) and also those that were least-predictive of 2-year TSA outcomes (subjective task-specific ADL questions). Future work should utilize this and related machine learning techniques to construct a more efficient and effective clinical outcome measure that incorporates subjective and objective input questions to better account for the preoperative factors that influence postoperative outcomes after TSA.

Funding

Exactech, Inc. (Gainesville, FL) provided support for clinical data collection but no funding to any author was provided to complete this study.

Disclaimers

Vikas Kumar, Steve Overman, and Ankur Teredesai are employed by Ken Sci, Inc. Christopher Roche is employed by Exactech, Inc. Ryan Simovitch and Howard Routman are consultants for Exactech, Inc. Pierre-Henri Flurin, Thomas Wright, and Joseph Zuckerman are consultants for Exactech, Inc. and receive royalties on products related to this article.

REFERENCES

[1] Amstutz HC, Sew Hoy AL, Clarke IC. UCLA anatomic total shoulder arthroplasty. Clin Orthop Relat Res 1981:(155):7–20.

[2] Roche C, Kumar V, Overman S, Simovitch R, Flurin PH, Wright T, et al. Shoulder Arthroplasty Smart Score. J Shoulder Elbow Surg. 2021. In press.

[3] Angst F, Goldhahn J, Pap G, Mannion AF, Roach KE, Siebertz D. et al. Cross-cultural adaptation, reliability and validity of the German Shoulder Pain and Disability Index (SPADI). Rheumatology (Oxford). 2007:(46):87–92. https://doi.org/10.1093/rheumatology/kel040.

[4] Angst F, Schwyzer HK, Aeschlimann A, Simmen BR, Goldhahn J. Measures of adult shoulder function: Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH) and its short version (QuickDASH), Shoulder Pain and Disability

Index (SPADI), American Shoulder and Elbow Surgeons (ASES) Society standardized shoulder assessment form, Constant (Murley) Score (CS), Simple Shoulder Test (SST), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire (SDQ), and Western Ontario Shoulder Instability Index (WOSI). Arthritis Care Res (Hoboken) 2011;63(Suppl 11):S174–88. https://doi.org/10.1002/acr.20630.
[5] Baumgarten KM, Chang PS. The American shoulder and elbow surgeons score highly correlates with the simple shoulder test [published online ahead of print, 2020 Jul 22]. J Shoulder Elbow Surg 2020;S1058-2746(20):30608–X. https://doi.org/10.1016/j.jse.2020.07.015.
[6] Beaton D, Richards RR. Assessing the reliability and responsiveness of 5 shoulder questionnaires. J Shoulder Elbow Surg 1998;7:565–72.
[7] Beaton DE, Richards RR. Measuring function of the shoulder. A cross-sectional comparison of five questionnaires. J Bone Joint Surg Am 1996;78:882–90.
[8] Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. J Rheumatol 2007;34:1426–31.
[9] Conboy VB, Morris RW, Kiss J, Carr AJ. An evaluation of the Constant-Murley shoulder assessment. J Bone Joint Surg Br 1996;78:229–32.
[10] Constant CR. Age related recovery of shoulder function after injury. Cork, Ireland: University College; 1986 MCh thesis.
[11] Constant CR, Murley AH. A clinical method of functional assessment of the shoulder. Clin Orthop Relat Res 1987:(214):160–4.
[12] Constant CR, Gerber C, Emery RJ, Søjbjerg JO, Gohlke F, Boileau P. A review of the constant score: modifications and guidelines for its use. J Shoulder Elbow Surg 2008;17:355–61. https://doi.org/10.1016/j.jse.2007.06.022.
[13] Cormack GV, Clarke CLA, Buttcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval; 2009. p. 758–9.
[14] Flurin PH, Roche CP, Wright TW, Marczuk Y, Zuckerman JD. A comparison and correlation of clinical outcome metrics in anatomic and reverse total shoulder arthroplasty. Bull Hosp Jt Dis 2013;73(Suppl 1):S118–23. 2015.
[15] 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.
[16] Goldhahn J, Angst F, Drerup S, Pap G, Simmen BR, Mannion AF. Lessons learned during the cross-cultural adaptation of the American Shoulder and Elbow Surgeons shoulder form into German. J Shoulder Elbow Surg 2008;17:248–54. https://doi.org/10.1016/j.jse.2007.06.027.
[17] 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.
[18] Hawkins RJ, Thigpen CA. Selection, implementation, and interpretation of patient-centered shoulder and elbow outcomes. J Shoulder Elbow Surg 2018;27:357–62. https://doi.org/10.1016/j.jse.2017.09.022.
[19] Hirschmann MT, Wind B, Amsler F, Gross T. Reliability of shoulder abduction strength measure for the Constant-Murley score. Clin Orthop Relat Res 2010;468(6):1565–71. https://doi.org/10.1007/s11999-009-1007-3.
[20] Jo YH, Lee KH, Jeong SY, Kim SJ, Lee BG. Shoulder outcome scoring systems have substantial ceiling effects 2 years after arthroscopic rotator cuff repair [published online ahead of print, 2020 May 21]. Knee Surg Sports Traumatol Arthrosc 2020;10. https://doi.org/10.1007/s00167-020-06036-y.
[21] Kocher MS, Horan MP, Briggs KK, Richardson TR, O'Holleran J, Hawkins RJ. Reliability, validity, and responsiveness of the American Shoulder and Elbow Surgeons subjective shoulder scale in patients with shoulder instability, rotator cuff disease, and glenohumeral arthritis. J Bone Joint Surg Am 2005;87:2006–11. https://doi.org/10.2106/JBJS.C.01624.
[22] 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. Clin Orthop Relat Res 2020;478 (10):2351–63. https://doi.org/10.1097/CORR.0000000000001263.
[23] Kumar V, Roche C, Overman S, Simovitch R, Flurin PH, Wright T. et al. Using machine learning to predict clinical outcomes after shoulder arthroplasty with a minimal feature set. J Shoulder Elbow Surg 2020 :S1058-2746(20)30646-7. doi: 10.1016/j.jse.2020.07.042.
[24] Michael RJ, Williams BA, Laguerre MD, Struk AM, Schoch BS, Wright TW. et al. Correlation of multiple patient-reported outcome measures across follow-up in patients undergoing primary shoulder arthroplasty. J Shoulder Elbow Surg 2019;28:1869–76. https://doi.org/10.1016/j.jse.2019.02.023.
[25] Michener LA, McClure PW, Sennett BJ. American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: reliability, validity, and responsiveness. J Shoulder Elbow Surg 2002;11:587–94. https://doi.org/10.1067/mse.2002.127096.
[26] Minoughan CE, Schumaier AP, Fritch JL, Grawe BM. Correlation of patient-reported outcome measurement information system physical function upper extremity computer adaptive testing, with the American Shoulder and Elbow Surgeons Shoulder Assessment Form and simple shoulder test in patients with shoulder pain. Arthroscopy 2018;34:1430–6. https://doi.org/10.1016/j.arthro.2017.11.040.
[27] Richards RR, An KN, Bigliani LU, Friedman RJ, Gartsman GM, Gristina AG. et al. A standardized method for the assessment of shoulder function. J Shoulder Elbow Surg 1994;3:347–52.
[28] Roche C, Simovitch R, Flurin PH, Wright T, Zuckerman J, Routman H. Comparison of the accuracy associated with three different machine learning models to predict outcomes after anatomic total shoulder arthroplasty and reverse total shoulder arthroplasty. Orthopaed Proc 2020;102-B(SUPP_1).
[29] Roy JS, MacDermid JC, Woodhouse LJ. A systematic review of the psychometric properties of the Constant-Murley score. J Shoulder Elbow Surg 2010;19:157–64. https://doi.org/10.1016/j.jse.2009.04.008.
[30] Sabesan VJ, Lombardo DJ, Khan J, Wiater JM. Assessment of the optimal shoulder outcome score for reverse shoulder arthroplasty. J Shoulder Elbow Surg 2015;24:1653–9. https://doi.org/10.1016/j.jse.2015.03.030.
[31] Sciascia AD, Morris BJ, Jacobs CA, Edwards TB. Responsiveness and internal validity of common patient-reported outcome measures following total shoulder arthroplasty. Orthopedics 2017;40:e513–9. https://doi.org/10.3928/01477447-20170327-02.
[32] Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27:623–56. 379−423.
[33] 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.
[34] Unger RZ, Burnham JM, Gammon L, Malempati CS, Jacobs CA, Makhni EC. The responsiveness of patient- reported outcome

tools in shoulder surgery is dependent on the underlying pathological condition. Am J Sports Med 2019;47:241–7. https://doi.org/10.1177/0363546517749213.

[35] Walton MJ, Walton JC, Honorez LA, Harding VF, Wallace WA. A comparison of methods for shoulder strength assessment and analysis of Constant score change in patients aged over fifty years in the United Kingdom. J Shoulder Elbow Surg 2007;16:285–9. https://doi.org/10.1016/j.jse.2006.08.002.

[36] Yian EH, Ramappa AJ, Arneberg O, Gerber C. The Constant score in normal shoulders. J Shoulder Elbow Surg 2005;14:128–33. https://doi.org/10.1016/j.jse.2004.07.003.

[37] Dowdle SB, Glass N, Anthony CA, Hettrich CM. Use of PROMIS for Patients Undergoing Primary Total Shoulder Arthroplasty. Orthop J Sports Med. 2017;5(9):2325967117726044. Published 2017 Sep 15. https://doi.org/10.1177/2325967117726044.