J Shoulder Elbow Surg (2021) ■, 1–14

ELSEVIER

Journal of Shoulder and Elbow Surgery

www.elsevier.com/locate/ymse

Validation of a machine learning–derived clinical metric to quantify outcomes after total shoulder arthroplasty

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

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

Background: We propose a new clinical assessment tool constructed using machine learning, called the Shoulder Arthroplasty Smart (SAS) score to quantify outcomes following total shoulder arthroplasty (TSA).

Methods: Clinical data from 3667 TSA patients with 8104 postoperative follow-up reports were used to quantify the psychometric properties of validity, responsiveness, and clinical interpretability for the proposed SAS score and each of the Simple Shoulder Test (SST), Constant, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form (ASES), University of California Los Angeles (UCLA), and Shoulder Pain and Disability Index (SPADI) scores.

Results: Convergent construct validity was demonstrated, with all 6 outcome measures being moderately to highly correlated preoperatively and highly correlated postoperatively when quantifying TSA outcomes. The SAS score was most correlated with the UCLA score and least correlated with the SST. No clinical outcome score exhibited significant floor effects preoperatively or postoperatively or significant ceiling effects preoperatively; however, significant ceiling effects occurred postoperatively for each of the SST (44.3%), UCLA (13.9%), ASES (18.7%), and SPADI (19.3%) measures. Ceiling effects were more pronounced for anatomic than reverse TSA, and generally, men, younger patients, and whites who received TSA were more likely to experience a ceiling effect than TSA patients who were female, older, and of non-white race or ethnicity. The SAS score had the least number of patients with floor and ceiling effects and also exhibited no response bias in any patient characteristic analyzed in this study. Regarding clinical interpretability, patient satisfaction anchor-based thresholds for minimal clinically importance difference and substantial clinical benefit were quantified for all 6 outcome measures; the SAS score thresholds were most similar in magnitude to the Constant score. Regarding responsiveness, all 6 outcome measures detected a large effect, with the UCLA exhibiting the most responsiveness and the SST exhibiting the least. Finally, each of the SAS, ASES, Constant, and SPADI scores had similarly large standardized response mean and effect size responsiveness.

Discussion: The 6-question SAS score is an efficient TSA-specific outcome measure with equivalent or better validity, responsiveness, and clinical interpretability as 5 other historical assessment tools. The SAS score has an appropriate response range without floor or

Institutional review board approval was received from all data collection sites.

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

1058-2746/$ - see front matter
https://doi.org/10.1016/j.jse.2021.01.021

ceiling effects and without bias in any target patient characteristic, unlike the age, gender, or race/ethnicity bias observed in the ceiling scores with the other outcome measures. Because of these substantial benefits, we recommend the use of the new SAS score for quantifying TSA outcomes.

Level of Evidence: Basic Science Study; Development and Validation of Outcome Instrument

Keywords: Patient-reported outcome measures; aTSA and rTSA outcomes; comparison psychometric properties; total shoulder arthroplasty; Shoulder Arthroplasty Smart score

Success after total shoulder arthroplasty (TSA) is determined by a combination of objective measures (ie, improvement of range of motion [ROM] in multiple planes, restoration of strength, and avoidance of complications) and subjective measures (ie, pain relief and functional recovery of activities of daily living). Therefore, any comprehensive assessment of TSA outcomes should consider both objective and subjective measures prior to surgery to provide a baseline and also assessments at various time points after surgery to quantify improvement achieved over time. Clinical outcome measures should also align with a patient's satisfaction with the procedure, which is informed by their perception of improvement relative to their expectations.

Patient-reported outcome measures (PROMs) are commonly used to quantify clinical outcomes. The advantage of these subjective assessment tools is that they are patient-centered, reflecting the patient's perspectives of both pretreatment status and treatment effectiveness. Consideration of these subjective measures is critical, as some treatment effects are only known to the patient, for example, pain intensity and pain relief. However, recent machine learning–based clinical research by Kumar et al[33] has demonstrated that not all subjective questions provide equivalent predictive validity. In particular, the task-specific activity of daily living questions used by the American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form (ASES) and Constant scores were found to be of low predictive value to postoperative TSA outcomes.[33] Alternatively, Kumar et al also demonstrated that both objective ROM measures and subjective measures of pain relief were highly predictive of postoperative TSA outcomes.

We propose and evaluate a new clinical assessment tool to quantify outcomes following anatomic TSA (aTSA) and reverse TSA (rTSA). This TSA-specific outcome measure aims to provide greater insight into both the disease progression and treatment by using both subjective and objective measures previously demonstrated[31-33] to be highly predictive of postoperative TSA outcomes. Doing so may create a new tool that better accounts for TSA outcomes variability. The goal of this study is to quantify and compare the psychometric properties of this machine learning derived assessment tool relative to 5 commonly used clinical measures to quantify TSA outcomes.

Methods

We analyzed a multicenter clinical outcomes database of shoulder arthroplasty patients who received a single-platform shoulder prosthesis (Equinoxe; Exactech, Inc., 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 30 different clinical sites according to an institutional review board–approved protocol. On completion of each form, all forms were independently verified and then computer scored on a secured IBM database. A total of 7947 shoulder arthroplasty patients were available for analysis. To ensure a homogenous data set of TSA patients, 1332 patients with revisions (n = 768), diagnosis of humeral fractures (n = 337), diagnosis of infections (n = 20), endoprostheses (n = 25), and hemiarthroplasty (n = 182) were excluded, leaving 6615 primary TSA patients available for analysis. To only analyze TSA patients after full recovery (ie, full improvement[49]), 2948 TSA patients with less than 2 years' follow-up were excluded, leaving 3667 primary aTSA and rTSA patients in the database with at least 2 years' follow-up for inclusion in this study.

The database contains demographic information, shoulder diagnoses, comorbidities, implant type, active and passive ROM, radiographic findings, and 5 clinical outcome scores: ASES, Constant, University of California Los Angeles (UCLA), Simple Shoulder Test (SST), and Shoulder Pain and Disability Index (SPADI), including the individual questions used to derive these 5 scores. All these data, including 291 preoperative inputs, were considered for item selection[28,29] in development of the new proposed clinical assessment tool. Regarding these 5 clinical outcome scores, it is important to note that some are PROMs as they only require patient feedback, whereas others like the Constant score and our proposed score are not true PROMs but instead are clinical outcome measures as they also require physical measurements and clinical input.

Shoulder Arthroplasty Smart Score

The proposed shoulder arthroplasty–specific clinical outcome measure, the Shoulder Arthroplasty Smart (SAS) score (see smartshoulderscore.com for an online calculator; Fig. 1), is a multidomain assessment consisting of 6 input questions, of which 3 are objective ROM measures and 3 are subjective measures of pain and function (Table I). The 6 input questions have an equal weight of 12.5 points each, and 1 additional input, called the Composite ROM score, is calculated from the 3 objective ROM measures to transform those values into a functional score with an allocated weight of 25 points, yielding a score range of 0-100 points, with 100 as the best score. These 6 questions were

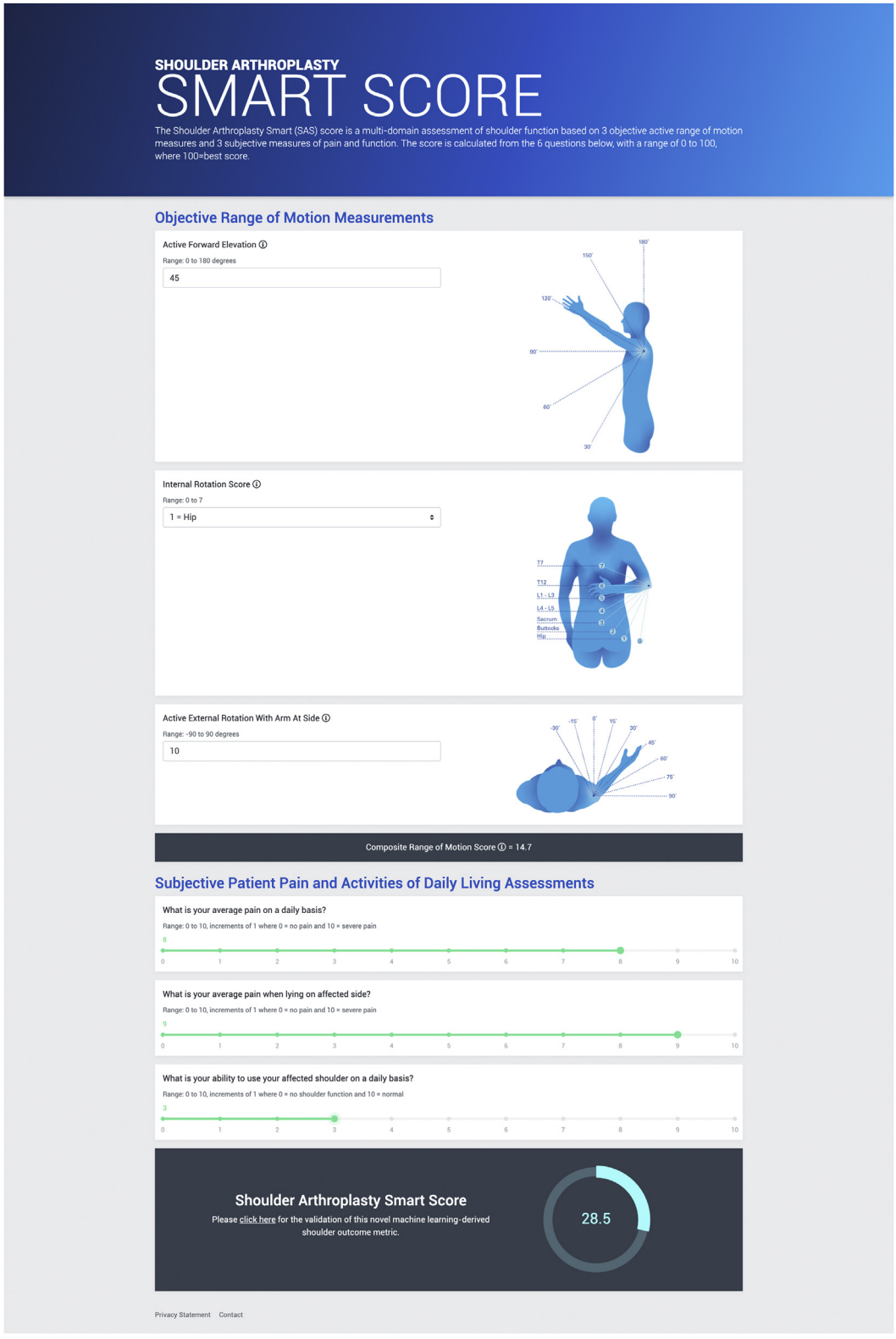

Figure 1 Smartshoulderscore.com website screenshot demonstrating graphical user interface of the online calculator.

Table I Input questions and scoring rationale for the Shoulder Arthroplasty Smart (SAS) score

Section	Question	Range	Scoring weight, %
Objective ROM measurement	Active forward elevation	0°-180°	12.5
Objective ROM measurement	Internal rotation score	0°-7°	12.5
Objective ROM measurement	Active external rotation with arm at side	–90° to 90°	12.5
Calculated ROM composite score	Transformation of 3 active ROM measurements into a functional composite ROM score (see Fig. 2)	0-100	25
Subjective pain	What is your average pain on a daily basis?	0-10 (0 = no pain, 10 = severe pain)	12.5
Subjective pain	What is your average pain when lying on affected side?	0-10 (0 = no pain, 10 = severe pain)	12.5
Subjective ADL function	What is your ability to use your affected shoulder on a daily basis?	0-10 (0 = no mobility, 10 = normal)	12.5
Total			100

ROM, range of motion; *ADL*, activity of daily living.
See smartshoulderscore.com for an online calculator.

identified using a machine learning analysis of this database and were selected because they are among the most predictive preoperative inputs that influence postoperative TSA outcomes.[31-33] Although the use of machine learning to identify and select the clinical tool's questions (from a larger data set consisting of 291 inputs) is new, it should be noted that the process to develop the SAS score is similar in structure to that recommended by Kirkley et al[28] and Kirshner et al,[29] except the work of a focus group to perform the item reduction is substituted by the machine learning analysis and identification of the most predictive preoperative inputs. As all 6 inputs already existed in the database, the SAS score was able to be retrospectively calculated from each patient's preoperative visit and all postoperative visits.

The objective ROM measures used by the SAS score are active forward elevation (range: 0°-180°), active external rotation with the arm at the side (range: –90° to 90°), and active internal rotation with the arm at the side, which is measured by an 8-point scale (range: 0-7) with the following discrete assignments based on motion to vertebral segments: 0° = 0, hip = 1, buttocks = 2, sacrum = 3, L5-L4 = 4, L3-L1 = 5, T12-T8 = 6, and T7 or higher = 7.[16] All ROM measures should be performed using a goniometer, as the baseline data at each clinical site was generated by this method. Previous machine learning research has demonstrated that the Composite ROM score is most predictive of TSA outcomes.[31,32] The Composite ROM score is calculated by the sum of 70% of the composite active forward elevation score, 15% of the composite active internal rotation score, and 15% of the active external rotation score, where these composite scores were adapted from the American Medical Association's Guides to the Evaluation of Permanent Impairment as graphically described in Figure 2.[1] The subjective measures used by the SAS score include 2 patient-assessed pain questions: (1) "What is your average pain on a daily basis?" (ie, visual analog scale pain score) and (2) "What is your average pain when lying on affected side?" (which was adopted from SPADI), each scored from 0-10, where 0 = no pain and 10 = severe pain; and a patient-assessed function question: "What is your ability to use your affected shoulder on a daily basis?" where 0 = no mobility and 10 = normal shoulder (ie, global shoulder function score).

This study quantifies the validity, responsiveness, and clinical interpretability of the SAS score and compares these psychometric properties to the 5 other outcome scores for aTSA and rTSA patients. Validity describes how accurately a clinical assessment tool quantifies what it is intending to measure. We quantify validity using descriptive statistics (including the mean, standard deviation, kurtosis, and skewness) and also a floor and ceiling effects analysis, which quantify the percentage of patients who experience the lowest and highest values of each score preoperatively and at each 2-year minimum postoperative follow-up visit. For construct validity, we perform a correlation analysis between all outcome measures to assess the consistency of the SAS score relative to its convergence with or divergence from the other 5 clinical outcome measures. Regarding these correlations between multiple groups, we used the false discovery rate–adjusted *P* value to control for multiple testing and type I error.

Responsiveness describes the sensitivity of a clinical assessment tool to detect a change. As described by Terwee et al, responsiveness is also a measure of longitudinal validity.[51] We quantify responsiveness of each clinical measure on the study population by using the effect size (ES) and the standardized response mean (SRM). ES is quantified by the mean pre- to postoperative improvement for a given measure relative to its preoperative standard deviation. As such, an ES of 1 equates to 1 standard deviation of change. Similarly, SRM is quantified by the mean pre- to postoperative improvement for a given measure relative to its pre- to postoperative standard deviation. The responsiveness of a clinical assessment tool should be greater for patients who experience a large change and smaller for those experiencing a small change. To demonstrate this, we quantify the ES and SRM for aTSA and rTSA patient subcohorts[20,30,52] while stratifying by a patient satisfaction item "Rate your satisfaction of treatment relative to your preoperative shoulder" with the responses "worse," "unchanged," "better," or "much better."

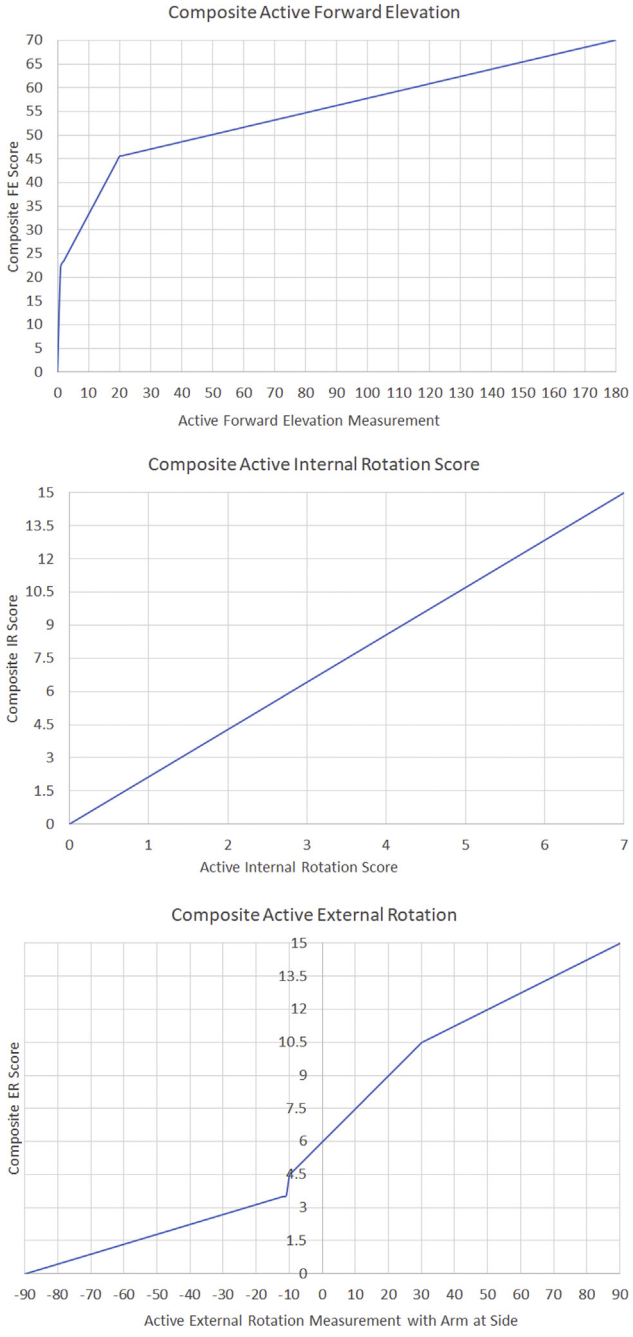

Figure 2 Composite range of motion transformation for active forward elevation (top), active internal rotation score (middle), and active external rotation with arm at side (bottom) measurements based on the American Medical Association Guides to the Evaluation of Permanent Impairment.[1]

Interpretability describes the clinical relevance of an outcome measures score. Clinical interpretability is most commonly assessed based on patient satisfaction, and in our study we quantify both the minimal clinically importance difference (MCID)[25,34,47] and the substantial clinical benefit (SCB)[19,39,48] using the aforementioned patient satisfaction anchor question to define thresholds of meaningful change for each outcome measure. Similar to the methodology used by Simovitch et al,[47,48] we quantify MCID[47] as the mean difference in improvement between TSA patients who described themselves as being "better" at each 2-year minimum follow-up visit as compared to patients who described themselves as "worse" and also "unchanged," and we quantify SCB[48] as the mean difference in improvement between "much better" patients as compared to "worse" and also "unchanged" patients.[47,48] As an alternative measure for MCID for each outcome measure, we also calculate MCID using the 0.5 distribution method as one-half of the standard deviation associated with the mean pre- to postoperative improvement.[41]

Results

Clinical data from 3667 patients (1594 aTSA, 2073 rTSA) with 8104 postoperative follow-up visits (3878 aTSA, 4226 rTSA) were used to quantify the psychometric properties for the SAS score and each of the SST, Constant, ASES, UCLA, and SPADI scores. (Table II) Notably, the SAS score had lower standard deviations preoperatively, postoperatively, and pre- to postoperatively for aTSA and rTSA cohorts as compared with the other 100-point scoring systems, demonstrating similar mean scores but less variance when measuring TSA outcomes. Table III demonstrates convergent construct validity with all 6 outcome measures being moderately to highly correlated preoperatively and highly correlated postoperatively when quantifying TSA outcomes. The SAS score was most correlated with the UCLA score (preoperative Pearson correlation coefficient $R = 0.78$, postoperative $R = 0.85$) and least correlated with the SST (preoperative $R = 0.63$, postoperative $R = 0.75$). Supplementary Tables S1–S3 present the kurtosis and skewness for all 6 outcome measures and demonstrate the SAS score was the most normally distributed measure preoperatively (ie, kurtosis closest to 0), the Constant score was the most normally distributed measure postoperatively, and the SST score was the most normally distributed measure for pre- to postoperative improvement, followed closely by the SAS score. All 6 outcome measures had highly negative skewness postoperatively and had moderately negative skewness in pre- to postoperative improvement, though the Constant and the SAS score distributions were the least-negatively skewed, suggesting a more normalized response range.

The pre- and postoperative floor (Table IV) and ceiling (Table V) effect analysis is presented in Tables IV and V. No clinical outcome score exhibited significant floor effects preoperatively or postoperatively; however, the strength component of the Constant score exhibited floor effects preoperatively in >60% of patients and postoperatively in >15% for both aTSA and rTSA. No clinical outcome score exhibited significant ceiling effects preoperatively; however, significant ceiling effects were present postoperatively for each of the SST (44.3%), UCLA (13.9%), ASES (18.7%), and SPADI (19.3%) measures. Ceiling effects were more pronounced for aTSA than rTSA, where each of

Table II Assessment of validity: comparison of preoperative, postoperative, and pre- to postoperative improvement outcomes between the 6 different clinical outcome measures

Preop/postop/pre- to postop improvement	aTSA + rTSA cohort (mean ± SD)	aTSA cohort (mean ± SD)	rTSA cohort (mean ± SD)
SST preop	3.8 ± 2.9	4.1 ± 3.0	3.6 ± 2.8
SST ≥2 yr postop	10.1 ± 2.5	10.5 ± 2.3	9.8 ± 2.7
SST pre- to postop improvement	6.4 ± 3.3	6.5 ± 3.3	6.3 ± 3.3
Constant preop	36.8 ± 14.1	38.5 ± 13.9	35.5 ± 14.1
Constant ≥2 yr postop	69.7 ± 14.3	71.6 ± 14.2	68.1 ± 14.1
Constant pre- to postop Improvement	33.5 ± 16.1	34.5 ± 15.8	32.8 ± 16.3
ASES preop	36.3 ± 16.1	36.3 ± 16.2	36.2 ± 16.0
ASES ≥2 yr postop	83.4 ± 18.7	85.0 ± 18.5	82.0 ± 18.8
ASES pre- to postop improvement	47.6 ± 21.7	49.7 ± 21.7	45.9 ± 21.6
UCLA preop	13.7 ± 4.1	14.3 ± 4.0	13.3 ± 4.2
UCLA ≥2 yr postop	30.4 ± 5.2	30.8 ± 5.3	30.0 ± 5.1
UCLA pre- to postop improvement	16.6 ± 5.9	16.7 ± 5.8	16.6 ± 6.0
SPADI preop	83.1 ± 23.2	82.4 ± 23.6	83.6 ± 22.9
SPADI ≥2 yr postop	20.6 ± 24.5	17.4 ± 22.7	23.5 ± 25.7
SPADI pre- to postop improvement	−62.6 ± 29.0	−65.7 ± 29.1	−60.0 ± 28.6
SAS preop	46.3 ± 11.5	46.4 ± 10.8	46.2 ± 12.1
SAS ≥2 yr postop	77.7 ± 12.2	80.4 ± 12.0	75.3 ± 11.9
SAS pre- to postop improvement	31.7 ± 14.6	34.9 ± 13.7	29.2 ± 14.8

SST, Simple Shoulder Test; *preop*, preoperation; *postop*, postoperation; *ASES*, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *UCLA*, University of California Los Angeles; *SPADI*, Shoulder Pain and Disability Index; *SAS*, Shoulder Arthroplasty Smart; *aTSA*, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty; *SD*, standard deviation.

Table III Assessment of validity: Correlation of preoperative and postoperative clinical outcome measure scores to each other and to patient satisfaction assessment

Combined aTSA + rTSA cohort	Patient satisfaction (preop/postop)	SST (preop/postop)	Constant (preop/postop)	ASES (preop/postop)	UCLA (preop/postop)	SPADI (preop/postop)
SST	0.096/0.435	1				
Constant	0.083/0.394	0.716/0.799	1			
ASES	0.075/0.460	0.717/0.841	0.653/0.810	1		
UCLA	0.064/0.521	0.630/0.756	0.773/0.806	0.771/0.873	1	
SPADI	−0.068/−0.444	−0.825/−0.890	−0.692/−0.805	−0.810/−0.924	−0.692/−0.823	1
SAS	0.074/0.418	0.630/0.748	0.781/0.847	0.694/0.832	0.783/0.852	−0.694/−0.825

aTSA, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty; *SST*, Simple Shoulder Test; *ASES*, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *UCLA*, University of California Los Angeles; *SPADI*, Shoulder Pain and Disability Index; *SAS*, Shoulder Arthroplasty Smart; *preop*, preoperation; *postop*, postoperation.
Pearson coefficient 0-0.3 is considered poor correlation, 0.3-0.6 is moderately correlated, 0.61-0.7 is strong to moderate, and >0.7 is highly correlated.[4]

the SST, UCLA, ASES, and SPADI measures had >20% of aTSA patients with ceiling effects. The SAS score had the least number of patients with floor and ceiling effects. A more detailed analysis demonstrated that male gender significantly increased the occurrence of a postoperative ceiling score for the SST ($P < .0001$), ASES ($P < .0001$), UCLA ($P = .0002$), and SPADI ($P < .0001$) measures (Table VI). Similarly, postoperative ceiling effects were more common in white than black patients for the SST ($P < .0001$), ASES ($P = .016$), UCLA ($P = .0014$), and SPADI ($P = .020$) measures; more common in white than Hispanic patients for the SST ($P = .0004$), ASES ($P = .016$), and SPADI ($P = .049$) measures; and more common in white than Asian patients for the SST ($P = .023$) (Table VII). Additionally, postoperative ceiling effects were observed between all 3 age groups for the SST, and for 2 of the 3 age groups for the ASES, UCLA, and SPADI measures (Table VIII).

The patient satisfaction anchor-based MCID (Table IX), the 0.5 distribution MCID (Table IX), and the patient

Table IV Assessment of validity: Comparison of floor effects across the 6 different clinical outcome measures

Clinical score	Reports with a "floor" score, %		
	aTSA + rTSA cohort (preop/postop)	aTSA cohort (preop/postop)	rTSA cohort (preop/postop)
SST	10.2/0.4	9.8/0.3	10.5/0.4
Constant	0.0/0.0	0.0/0.0	0.0/0.0
Strength component of Constant score	67.0/15.9	61.0/15.4	72.0/16.3
ASES	0.1/0.0	0.1/0.0	0.1/0.0
UCLA	0.0/0.0	0.0/0.0	0.0/0.0
SPADI	0.6/0.1	0.6/0.0	0.5/0.1
SAS	0.0/0.0	0.0/0.0	0.0/0.0

SST, Simple Shoulder Test; *ASES*, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *UCLA*, University of California Los Angeles; *SPADI*, Shoulder Pain and Disability Index; *SAS*, Shoulder Arthroplasty Smart; *aTSA*, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty; *preop*, preoperation; *postop*, postoperation.

Table V Assessment of validity: Comparison of ceiling effects across the 6 different clinical outcome measures

Clinical score	Reports with a "ceiling" score, %		
	aTSA + rTSA cohort (preop/postop)	aTSA cohort (preop/postop)	rTSA cohort (preop/postop)
SST	1.0/44.3	1.1/52.9	1.0/36.7
Constant	0.0/0.3	0.0/0.5	0.0/0.1
Strength component of Constant score	0.2/0.7	0.5/1.1	0.0/0.4
ASES	0.0/18.7	0.0/25.1	0.0/12.8
UCLA	0.0/13.9	0.0/20.8	0.0/7.9
SPADI	0.0/19.3	0.0/25.6	0.0/13.6
SAS	0.0/0.1	0.0/0.2	0.0/0.0

SST, Simple Shoulder Test; *ASES*, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *UCLA*, University of California Los Angeles; *SPADI*, Shoulder Pain and Disability Index; *SAS*, Shoulder Arthroplasty Smart; *aTSA*, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty; *preop*, preoperation; *postop*, postoperation.

satisfaction anchor-based SCB (Table X) thresholds for all 6 outcome measures is presented in Tables IX and X. The anchor-based MCID and SCB thresholds for aTSA patients were larger than the MCID and SCB values for rTSA patients for all outcome measures, except the SST. However, no differences were observed between aTSA and rTSA patients for the 0.5 distribution MCID thresholds. Finally, the MCID and SCB thresholds for the SAS score were most similar in magnitude to the Constant score.

The SRM and ES responsiveness for all 6 outcome measures is presented in Table XI and for the combined cohort of aTSA and rTSA when stratified by patient satisfaction ranking in Supplementary Table S4. At 2 years' minimum follow-up, all 6 clinical outcome measures detected a large effect (as defined by an SRM or ES >0.8),[9,15,35] with the UCLA exhibiting the most responsiveness and the SST exhibiting the least; each of the SAS, ASES, Constant, and SPADI scores had similarly large SRM and ES responsiveness. Finally, all 6 outcome measures demonstrated a greater response for patients who experienced a large change and a smaller response for patients who experienced a small change, as noted by the stepped SRM and ES values for each of the patient satisfaction rankings. However, as described in Supplementary Table S4, even patients who reported their shoulder as "worse" still exhibited mean pre- to postoperative improvement.

Discussion

We present the first orthopedic clinical outcome measure derived using machine learning and constructed of preoperative inputs that are most predictive of postoperative TSA outcomes. The results of this 3667–TSA outcome study demonstrate that the SAS score has equivalent or better psychometric properties of validity, responsiveness, and clinical interpretability as the ASES, Constant, UCLA, SST, and SPADI scores. The SAS score has an appropriate response range with no floor or ceiling effects for aTSA or

Table VI Assessment of validity: Comparison of ceiling effects on gender across the 6 different clinical outcome measures

Clinical score	Postoperative reports with a "ceiling" score		
	aTSA + rTSA cohort (male), %	aTSA + rTSA cohort (female), %	*P* value (male vs. female)
SST	56.4	35.7	**<.0001**
Constant	0.4	0.2	NS
ASES	23.8	14.9	**<.0001**
UCLA	15.8	12.6	**.0002**
SPADI	22.5	17.0	**<.0001**
SAS	0.1	0.1	NS

SST, Simple Shoulder Test; *ASES*, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *UCLA*, University of California Los Angeles; *SPADI*, Shoulder Pain and Disability Index; *SAS*, Shoulder Arthroplasty Smart; *aTSA*, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty; *NS*, not significant.
Boldface indicates significance ($P < .05$).

rTSA and no gender, age, or race or ethnicity response bias, as was observed with many of the other historical scores when quantifying TSA outcomes. These psychometric improvements were achieved despite the SAS score consisting of only 6 input questions; this efficient selection of 3 objective and 3 subjective measures is useful for both quality assurance and clinical research purposes, and represents a reduction of approximately half (or more) of the inputs required by ASES, SST, SPADI, and Constant scores. Such efficiency will likely reduce administrative burden and responder fatigue while improving patient compliance when performing TSA clinical research. By combining subjective and objective measures, SAS remains patient focused in order to avoid perception pitfalls between patient and physician on subjective measures[28,50] while also using physical measurements that objectively quantify functional capability and ROM limitations. Of note, 6 input questions with SAS is similar in range to the 4-6 questions reportedly used by the Patient Reported Outcomes Measurement Information System Upper Extremity (PROMIS UE) computer adaptive test,[7,14] suggesting that an appropriate selection of only the most predictive input questions can eliminate floor or ceiling effects without the need for an adaptive algorithm and a large hierarchically structured item bank, like the 46 questions used by the PROMIS UE.[40] Future work should compare the psychometric properties of the SAS score to other next-generation PROMs and clinical measures, like the PROMIS.

A clinical outcome measure should offer insight into both the disease progression and treatment for an individual patient; as such, a clinical outcome measures must be valid for both pre- and postoperative assessments. Numerous studies have analyzed the preoperative floor or ceiling effects of the clinical outcome measures analyzed in this study[18,20,21,30,40]; however, only a few have analyzed the postoperative floor or ceiling effects,[26,46] which are equally important and necessary to demonstrate that a tool provides an appropriate response range without a response bias. Our postoperative ceiling findings for TSA patients using the SST (44.3%), UCLA (13.9%), ASES (18.7%), SPADI (19.3%), and Constant (0.3%) scores are similar and in range of those reported by Jo et al[26] (28% ASES, 34% SST, 30% UCLA, and 20% Constant) and Sciascia et al[46] (21% ASES and 3% Constant). To be internally valid, a clinical outcome measure should not exhibit floor or ceiling effects in >15% of patients,[35,36] as the clinical assessment tool is insensitive to detect change in the negative direction for patients with a floor score or in the positive direction for ceiling scores. Therefore, the presence of any floor or ceiling effect negatively impacts responsiveness and longitudinal validity of an outcome measure. The results of this 2-year minimum TSA outcome study demonstrate that the strength component of the Constant score exhibits floor effects preoperatively in >60% of aTSA and rTSA patients and postoperatively in >15% of aTSA and rTSA patients. Additionally, each of the SST, UCLA, ASES, and SPADI scores exhibit postoperative ceiling effects in >15% of patients, particularly for aTSA patients. Generally, patients who were male, younger, white, and had received TSA were more likely to experience a ceiling effect, as compared to patients who were female, older, and of nonwhite race or ethnicity. These findings are also reflected by the negative skewness of each outcome measure and suggest that all 5 historical clinical measures have an insufficient response range to quantify TSA outcomes and the finding of different postoperative ceiling effects for different patient cohorts of age, gender, and race or ethnicity, suggests that each of the SST, UCLA, ASES, and SPADI scores also have a response bias when quantifying TSA outcomes. Furthermore, the high occurrence of postoperative ceiling effects questions the efficacy of these historical tools for longitudinal evaluation, given their limited ability to accurately quantify outcomes with ceiling score patients in

Table VII Assessment of validity: Comparison of ceiling effects on race or ethnicity across the 6 different clinical outcome measures

Clinical score	Postoperative reports with a "ceiling" score				
	aTSA + rTSA cohort (white [W]), %	aTSA + rTSA cohort (black [B]), %	aTSA + rTSA cohort (Hispanic [H]), %	aTSA + rTSA cohort (Asian [A]), %	*P* value after FDR multiple test correction
SST	45.9	29.3	27.4	20.7	W vs. B: **<.0001** W vs. H: **.0004** B vs. H: NS W vs. A: **.023** B vs. A: NS H vs. A: NS
Constant	0.3	0.0	0.0	0.0	W vs. B: NS W vs. H: NS B vs. H: NS W vs. A: NS B vs. A: NS H vs. A: NS
ASES	19.8	12.4	8.8	6.3	W vs. B: **.016** W vs. H: **.016** B vs. H: NS W vs. A: NS B vs. A: NS H vs. A: NS
UCLA	14.3	5.4	13.8	10.7	W vs. B: **.0014** W vs. H: NS B vs. H: NS W vs. A: NS B vs. A: NS H vs. A: NS
SPADI	20.3	12.6	10.7	9.7	W vs. B: **.020** W vs. H: **.049** B vs. H: NS W vs. A: NS B vs. A: NS H vs. A: NS
SAS	0.1	0.0	0.0	0.0	W vs. B: NS W vs. H: NS B vs. H: NS W vs. A: NS B vs. A: NS H vs. A: NS

SST, Simple Shoulder Test; *ASES*, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *UCLA*, University of California Los Angeles; *SPADI*, Shoulder Pain and Disability Index; *SAS*, Shoulder Arthroplasty Smart; *aTSA*, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty; FDR, false discovery rate; *NS*, not significant.
Boldface indicates significance ($P < .05$).

this 2-year minimum follow-up study, and also suggests that these historical tools are only useful longitudinally for patients with severe dysfunction.[36]

To be truly valid, a clinical outcome measure must not give rise to bias across the unique characteristics of the target patient population in which it is used, such that no variations in results will occur from use of the tool. Importantly, no bias was observed for any studied patient characteristic with the SAS score when quantifying TSA outcomes. The finding of bias with these commonly used and historical clinical outcome tools in this TSA patient population is concerning and is itself a new finding for each of the ASES, SST, UCLA, and SPADI scores, despite each score being previously "validated." The age and gender bias present in the Constant score is inherent, as the Constant score defines a normal shoulder as that of a 25-year-old man.[10-12] Floor effects with the Constant strength assessment have been previously reported[8] and are primarily due to the measurement being performed at 90° abduction, as some TSA patients are unable to achieve this position both before or after surgery,[8,24,42] let alone hold a weight-in-hand (up to 11.3 kg). As a result, it is a common

Table VIII Assessment of validity: Comparison of ceiling effects on patient age at the time of surgery across the 6 different clinical outcome measures

Clinical score	Postoperative reports with a "ceiling" score			
	aTSA + rTSA cohort (<60 yr), %	aTSA + rTSA cohort (60-79 yr), %	aTSA + rTSA cohort (≥80 yr), %	*P* value after FDR multiple test correction
SST	42.7	46.4	30.0	<60 vs. 60-79 yr: **.038** <60 vs. ≥80 yr: **<.0001** 60-79 vs. ≥80 yr: **<.0001**
Constant	0.7	0.2	0.2	<60 vs. 60-79 yr: NS <60 vs. ≥80 yr: NS 60-79 vs. ≥80 yr: NS
ASES	17.9	19.5	13.0	<60 vs. 60-79 yr: NS <60 vs. ≥80 yr: **.009** 60-79 vs. ≥80 yr: **<.0001**
UCLA	15.7	14.4	7.6	<60 vs. 60-79 yr: NS <60 vs. ≥80 yr: **<.0001** 60-79 vs. ≥80 yr: **<.0001**
SPADI	17.0	20.4	13.8	<60 vs. 60-79 yr: **.024** <60 vs. ≥80 yr: NS 60-79 vs. ≥80 yr: **<.0001**
SAS	0.0	0.1	0.1	<60 vs. 60-79 yr: NS <60 vs. ≥80 yr: NS 60-79 vs. ≥80 yr: NS

SST, Simple Shoulder Test; *ASES*, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *UCLA*, University of California Los Angeles; *SPADI*, Shoulder Pain and Disability Index; *SAS*, Shoulder Arthroplasty Smart; *aTSA*, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty; *NS*, not significant.
Boldface indicates significance ($P < .05$).

Table IX Assessment of responsiveness and interpretability: Comparison of patient satisfaction anchor-based and 0.5 SD distribution-based MCID across the 6 different clinical outcome measures

Clinical score	Anchor MCID/ distribution MCID		
	aTSA + rTSA cohort	aTSA cohort	rTSA cohort
SST	1.8/1.7	1.7/1.7	1.8/1.7
Constant	5.3/8.0	8.6/7.9	3.0/8.1
ASES	12.4/10.9	14.2/10.8	11.2/10.8
UCLA	7.9/3.0	8.1/2.9	7.7/3.0
SPADI	−20.4/−14.5	−19.7/−14.6	−21.3/−14.3
SAS	6.1/7.3	8.5/6.9	4.9/7.4

SD, standard deviation; *MCID*, minimal clinically important difference; *SST*, Simple Shoulder Test; *ASES*, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *UCLA*, University of California Los Angeles; *SPADI*, Shoulder Pain and Disability Index; *SAS*, Shoulder Arthroplasty Smart; *aTSA*, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty.

practice to perform an age and gender normalization for the Constant score when quantifying TSA outcomes. However, this normalization is a source of error as it transforms a patient-specific score using an assumed average decline for age and gender.[10,12,27,54,55] Yian et al[55] demonstrated that not all patients decline at the same rate and the normalization method originally presented by Constant[10,12] is not representative of, or generalizable to, all patients. Specifically, Constant's recommended normalization method may overestimate outcomes for female patients >40 years old and male patients >60 years old.[55] In our study, 99.7% of women were >40 years old and 82.4% of men were >60 years old at the time of their TSA; because of this finding, we report the Constant score as an absolute score rather than a normalized score.

In the absence of a gold standard TSA outcome measure, validity should be considered on a continuum, with increasing confidence in that analysis based on the size of the study and resulting clinical evidence.[17] This psychometric analysis of 6 different outcome measures used data

Table X Assessment of responsiveness and interpretability: Comparison of patient satisfaction anchor-based SCB across the 6 different clinical outcome measures

Clinical score	SCB		
	aTSA + rTSA cohort	aTSA cohort	rTSA cohort
SST	3.5	3.5	3.5
Constant	16.9	20.4	14.3
ASES	30.7	33.2	28.7
UCLA	11.8	12.6	11.3
SPADI	−44.0	−44.3	−43.7
SAS	16.6	19.2	14.4

SCB, substantial clinical benefit; *SST*, Simple Shoulder Test; *ASES*, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *UCLA*, University of California Los Angeles; *SPADI*, Shoulder Pain and Disability Index; *SAS*, Shoulder Arthroplasty Smart; *aTSA*, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty.

from 3667 TSA patients with 8104 postoperative follow-up visits; as such, it is by far the largest study of its kind to assess the validity, responsiveness, and clinical interpretability of any TSA outcome measures. Furthermore, given that these large-scale clinical data were generated from 30 different clinical sites and surgeons in both the United States and Europe, the results of our study are more generalizable and substantially broader than the evidence that can be contributed by a single site or surgeon. Our findings demonstrate that the SAS score has equivalent or better validity and responsiveness than the SST, Constant, ASES, UCLA, and SPADI scores when quantifying TSA outcomes. Our findings of validity and responsiveness with these 5 historical outcome measures were in the range of previously published studies.[2-6,13,16,20,26,30,37,38,45,46,53] Regarding clinical interpretability, the MCID and SCB thresholds for the SAS score closely match those of the Constant score, which is also a 100-point system with a low occurrence of ceiling effects. Additionally, our findings of MCID and SCB thresholds for these 5 historical outcome measures closely match the findings reported by Simovitch et al.[47,48]

Active forward elevation features prominently in the SAS score calculation, accounting for 30% of the overall score (ie, 12.5% contribution as 1 of the 6 input questions + 70% contribution of the Composite ROM score). Active forward elevation is critical for patients to perform many activities of daily living, with substantial impairment occurring when forward elevation is <20°.[1] Goodman et al[22] recently reported that small improvements in forward elevation can improve overall health, demonstrating significant correlations with both the physical and mental component of the 12-Item Short Form Health Survey (SF-12). Importantly, Goodman et al also reported that forward elevation was the only measure considered by the ASES physician assessment that demonstrated any significant impact on clinical improvement for rTSA patients. These findings support the emphasis of active forward elevation in the SAS score calculation, and beyond that, it reinforces the need for a new clinical outcome measure composed of inputs more predictive of postoperative TSA outcomes. However, we recognize that restoration of ROM may not be the most important consideration for all TSA patients, and for those patients the SAS score calculation may overemphasize functional improvement. Additionally, not all clinics may use a goniometer for ROM assessment, and for those clinics this score may not be compatible with their existing workflow.

This study has several limitations. First, we did not perform a reliability assessment, which is an essential component of a psychometric analysis. However, it is important to note that the SAS score is constructed from questions composed of other scores that were previously validated for reliability,[2,3,5,21,30,38,43,44] and thus, it is reasonable to assume the reliability of those questions should be inherited. Second, this internal validation study was performed retrospectively on the same data set in which the machine learning analysis was conducted to

Table XI Assessment of responsiveness: Comparison of SRM and ES for the overall cohort across the 6 different clinical outcome measures

Clinical score	SRM/ES for overall cohort		
	aTSA + rTSA cohort	aTSA cohort	rTSA cohort
SST	1.91/2.20	1.96/2.19	1.88/2.23
Constant	2.09/2.38	2.18/2.49	2.02/2.32
ASES	2.19/2.95	2.29/3.06	2.12/2.87
UCLA	2.81/4.03	2.88/4.17	2.76/3.97
SPADI	−2.16/−2.70	−2.26/−2.79	−2.10/−2.62
SAS	2.17/2.75	2.55/3.24	1.97/2.41

SRM, standardized response mean; *ES*, effect size; *SST*, Simple Shoulder Test; *ASES*, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *UCLA*, University of California Los Angeles; *SPADI*, Shoulder Pain and Disability Index; *SAS*, Shoulder Arthroplasty Smart; *aTSA*, anatomic total shoulder arthroplasty; *rTSA*, reverse total shoulder arthroplasty.

identify and select the SAS questions; thus, future work is necessary to externally validate our findings. Third, we demonstrated age, gender, and race or ethnicity bias with multiple historical clinical measures for postoperative TSA outcomes; however, it should be recognized that the number of patients in each subcohort was unequal, as the patients comprising our database are predominately elderly and white. Fourth, as there is no gold standard assessment tool to quantify TSA outcomes,[3,23] we were unable to quantify criterion validity for the SAS score. Instead, we compared pre- and postoperative SAS score results to 5 commonly used outcome measures and demonstrated convergent construct validity. Fifth, whereas the inputs of the SAS score were constructed using machine learning,[31-33] a SAS predictive model has not yet been developed. Future work should create a predictive model for the SAS score to maximize the utility of this new assessment tool for the clinical researcher. Sixth, although the SAS tool was constructed using multidimensional data, including 5 different PROMs and clinical outcome measures and their associated input questions, it is likely that there are other data and also other scores that may be more predictive than what we considered in our machine learning analysis. Thus, future research efforts should strive to obtain the most relevant clinical data for TSA patients, and on collection, this machine learning item selection process should be repeated. Finally, recent[23] clinical research guidelines recommend combining a shoulder-specific outcome measure with a generic quality of life measure to establish a baseline assessment of health beyond comorbidities.[23] We agree with this recommendation as all shoulder clinical tools generally only represent health as the absence of functional shoulder limitation[36]; however, we did not use any quality of life assessment tool in the item selection[28,29] or item reduction[28,29] process to develop the SAS score.

Conclusion

The SAS score (smartshoulderscore.com) is the first orthopedic clinical outcome measure constructed using machine learning. Our psychometric analysis of 3667 TSA patients demonstrates that this new assessment tool has equivalent or better validity, responsiveness, and clinical interpretability as 5 other measures to quantify TSA outcomes. Consisting of only 6 inputs, the SAS score represents an efficiency improvement of half (or more) the number of input questions relative to these other measures. Additionally, our psychometric analysis demonstrates that the SAS score has an appropriate response range without floor or ceiling effects and without bias in any target patient characteristic, unlike the age, gender, or race or ethnicity bias observed in the ceiling scores for the historical measures analyzed in this study. Because of these substantial benefits, we recommend the use of the new SAS score for quantifying TSA outcomes, though future work remains to perform external validations and quantify the reliability of use of this machine learning–based outcome tool.

Acknowledgments

We would like to express our appreciation to John David Eriksen, Daniel Holth, and Jenna Spencer for developing the smartshoulderscore.com online calculator and to Wen Fan for her statistical expertise and analysis.

Disclaimer

Exactech, Inc. (Gainesville, FL, USA) funded data collection at all site; however, no author was paid to conduct this study or write this paper.

Christopher Roche is employed by Exactech, Inc.

Vikas Kumar, Steve Overman, and Ankur Teredesai are employed by Ken Sci, 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.

Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jse.2021.01.021.

References

1. American Medical Association. Guides to the evaluation of permanent impairment. In: Rondinelli R, editor. 6th edition. Chicago, IL: American Medical Association; 2008.
2. 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
3. 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
4. Baumgarten KM, Chang PS. The American Shoulder and Elbow Surgeons score highly correlates with the Simple Shoulder Test. J

Shoulder Elbow Surg 2021;30:707-11. https://doi.org/10.1016/j.jse.2020.07.015.
5. Beaton D, Richards RR. Assessing the reliability and responsiveness of 5 shoulder questionnaires. J Shoulder Elbow Surg 1998;7:565-72.
6. 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.
7. Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. J Rheumatol 2007;34:1426-31.
8. Christie A, Hagen KB, Mowinckel P, Dagfinrud H. Methodological properties of six shoulder disability measures in patients with rheumatic diseases referred for shoulder surgery. J Shoulder Elbow Surg 2009;18:89-95. https://doi.org/10.1016/j.jse.2008.07.008
9. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. New York: Lawrence Erlbaum Associates; 1998.
10. Constant CR. Age related recovery of shoulder function after injury. MCh thesis. Cork, Ireland: University College; 1986.
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. Cronin KJ, Magnuson JA, Murphy ML, Unger RZ, Jacobs CA, Blake MH. Responsiveness of patient reported outcomes in shoulder arthroplasty: what are we actually measuring? J Shoulder Elbow Surg 2020. Epub ahead of print. https://doi.org/10.1016/j.jse.2020.08.019
14. Dowdle SB, Glass N, Anthony CA, Hettrich CM. Use of PROMIS for patients undergoing primary total shoulder arthroplasty. Orthop J Sports Med 2017;5:2325967117726044. https://doi.org/10.1177/2325967117726044
15. Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. Meta-analysis of experiments with matched groups or repeated measures designs. Psychol Methods 1996;1:170.
16. 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) 2015;73(suppl 1):S118-23.
17. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD, Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health 2007;10(suppl 2):S94-105. https://doi.org/10.1111/j.1524-4733.2007.00272.x
18. Fu MC, Chang B, Wong AC, Nwachukwu BU, Warren RF, Dines DM, et al. PROMIS physical function underperforms psychometrically relative to American Shoulder and Elbow Surgeons score in patients undergoing anatomic total shoulder arthroplasty. J Shoulder Elbow Surg 2019;28:1809-15. https://doi.org/10.1016/j.jse.2019.02.011
19. Glassman SD, Copay AG, Berven SH, Polly DW, Subach BR, Carreon LY. Defining substantial clinical benefit following lumbar spine arthrodesis. J Bone Joint Surg Am 2008;90:1839-47. https://doi.org/10.2106/JBJS.G.01095
20. Godfrey J, Hamman R, Lowenstein S, Briggs K, Kocher M. Reliability, validity, and responsiveness of the simple shoulder test: psychometric properties by age and injury type. J Shoulder Elbow Surg 2007;16:260-7. https://doi.org/10.1016/j.jse.2006.07.003
21. 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
22. Goodman J, Lau BC, Krupp RJ, Getz CL, Feeley BT, Ma CB, et al. Clinical measurements versus patient-reported outcomes: analysis of the American Shoulder and Elbow Surgeons physician assessment in patients undergoing reverse total shoulder arthroplasty. JSES Open Access 2018;2:144-9. https://doi.org/10.1016/j.jses.2018.01.003
23. 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
24. Hirschmann MT, Wind B, Amsler F, Gross T. Reliability of shoulder abduction strength measure for the Constant-Murley score. Clin Orthop Relat Res 2010;468:1565-71. https://doi.org/10.1007/s11999-009-1007-3
25. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407-15.
26. 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 [Epub ahead of print]. Knee Surg Sports Traumatol Arthrosc 2020. https://doi.org/10.1007/s00167-020-06036-y
27. Katolik LI, Romeo AA, Cole BJ, Verma NN, Hayden JK, Bach BR. Normalization of the Constant score. J Shoulder Elbow Surg 2005;14:279-85. https://doi.org/10.1016/j.jse.2004.10.009
28. Kirkley A, Griffin S, McLintock H, Ng L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI). Am J Sports Med 1998;26:764-72.
29. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis 1985;38:27-36.
30. 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
31. 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;478:2351-63. https://doi.org/10.1097/CORR.0000000000001263
32. 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. Epub ahead of print. https://doi.org/10.1016/j.jse.2020.07.042
33. Kumar V, Roche C, Overman S, Simovitch R, Flurin PH, Wright T, et al. Use of machine learning to assess the predictive value of 3 commonly used clinical measures to quantify outcomes after total shoulder arthroplasty [Epub ahead of print]. Semin Arthroplasty JSES 2020. https://doi.org/10.1053/j.sart.2020.12.003
34. Leopold SS, Porcher R. Editorial: The minimum clinically important difference—the least we can do. Clin Orthop Relat Res 2017;475:929-32. https://doi.org/10.1007/s11999-017-5253-5
35. Lohr KN, Aaronson NK, Alonso J, Burtnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther 1996;18:979-92.
36. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293-307.
37. 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
38. 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
39. Michener LA, Snyder Valier AR, McClure PW. Defining substantial clinical benefit for patient-rated outcome tools for shoulder

impingement syndrome. Arch Phys Med Rehabil 2013;94:725-30. https://doi.org/10.1016/j.apmr.2012.11.011
40. 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
41. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41:582-92. https://doi.org/10.1097/01.MLR.0000062554.74615.4C
42. Othman A, Taylor G. Is the Constant score reliable in assessing patients with frozen shoulder? 60 shoulders scored 3 years after manipulation under anaesthesia. Acta Orthop Scand 2004;75:114-6. https://doi.org/10.1080/00016470410001708230
43. 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.
44. Roach KE, Budiman-Mak E, Songsiridej N, Lertratanakul Y. Development of a shoulder pain and disability index. Arthritis Care Res 1991;4:143-9.
45. 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
46. 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
47. 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
48. 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
49. 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
50. Sprangers MA, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. J Clin Epidemiol 1992;45:743-60.
51. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42. https://doi.org/10.1016/j.jclinepi.2006.03.012
52. Thigpen CA, Shanley E, Momaya AM, Kissenberth MJ, Tolan SJ, Tokish JM, et al. Validity and responsiveness of the single alphanumeric evaluation for shoulder patients. Am J Sports Med 2018;46:3480-5. https://doi.org/10.1177/0363546518807924
53. 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
54. 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
55. 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