J Shoulder Elbow Surg (2021) 30, e50–e59

ELSEVIER

Journal of Shoulder and Elbow Surgery

www.elsevier.com/locate/ymse

Comparison of machine learning techniques to predict unplanned readmission following total shoulder arthroplasty

Varun Arvind, BS, Daniel A. London, MD, MS, Carl Cirino, MD, Aakash Keswani, MD, Paul J. Cagle, MD*

Department of Orthopaedic Surgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Background: Machine learning (ML) techniques have been shown to successfully predict postoperative complications for high-volume orthopedic procedures such as hip and knee arthroplasty and to stratify patients for risk-adjusted bundled payments. The latter has not been done for more heterogeneous, lower-volume procedures such as total shoulder arthroplasty (TSA) with equally limited discussion around strategies to optimize the predictive ability of ML algorithms. The purpose of this study was to (1) assess which of 5 ML algorithms best predicts 30-day readmission, (2) test select ML strategies to optimize the algorithms, and (3) report on which patient variables contribute most to risk prediction in TSA across algorithms.

Methods: We identified 9043 patients in the American College of Surgeons National Surgical Quality Improvement Database who underwent primary TSA between 2011 and 2015. Predictors included demographics, comorbidities, laboratory data, and intraoperative variables. The outcome of interest was 30-day unplanned readmission. Five ML algorithms—support-vector machine (SVM), logistic regression, random forest (RF), an adaptive boosting algorithm, and neural network—were trained on the derivation cohort (2011-2014 TSA patients) to predict 30-day unplanned readmission rates. After training, weights for each respective model were fixed and the classifiers were evaluated on the 2015 TSA cohort to simulate a prospective evaluation. C-statistic and f1 scores were used to assess the performance of each classifier. After evaluation, features were removed independently to assess which features most affected classifier performance.

Results: The derivation and validation cohorts comprised 5857 and 3186 primary TSA patients, respectively, with similar demographics, comorbidities, and 30-day unplanned readmission rates (2.9% vs. 2.7%). Of the ML algorithms, SVM performed the worst with a c-statistic of 0.54 and an f1-score of 0.07, whereas the random-forest classifier performed the best with the highest c-statistic of 0.74 and an f1-score of 0.18. In addition, SVM was most sensitive to loss of single features, whereas the performance of RF did not dramatically decrease after loss of single features. Within the trained RF classifier, 5 variables achieved weights >0.5 in descending order: high bilirubin (>1.9 mg/dL), age >65, race, chronic obstructive pulmonary disease, and American Society of Anesthesiologists' scores ≥ 3. In our validation cohort, we observed a 2.7% readmission rate. From this cohort, using the RF classifier we were then able to identify 436 high-risk patients with a predicted risk score >0.6, of whom 36 were readmitted (readmission rate of 8.2%).

Conclusion: Predictive analytics algorithms can achieve acceptable prediction of unplanned readmission for TSA with the RF classifier outperforming other common algorithms.

Level of evidence: Basic Science Study; Computer Modeling

Published by Elsevier Inc. on behalf of Journal of Shoulder and Elbow Surgery Board of Trustees.

Keywords: Machine learning; total shoulder arthroplasty; readmission; bundled payments; risk stratification; risk assessment

This basic science study was exempt from institutional review board approval.

*Reprint requests: Paul J. Cagle, MD, Icahn School of Medicine at Mount Sinai, 425 West 59th Street, 5th Floor, New York, NY 10029, USA.

E-mail address: paul.cagle@mountsinai.org (P.J. Cagle).

1058-2746/$ - see front matter Published by Elsevier Inc. on behalf of Journal of Shoulder and Elbow Surgery Board of Trustees.
https://doi.org/10.1016/j.jse.2020.05.013

Total shoulder arthroplasty (TSA) is a commonly performed surgery for the treatment of end-stage glenohumeral arthritis. The current procedural terminology for TSA also includes reverse TSA, which is indicated for rotator cuff arthropathy, massive irreparable rotator cuff tears, and complex 3- and 4-part proximal humerus fractures among others. Since FDA approval of reverse TSA in 2004, the prevalence of TSA has increased nearly 3-fold over the past decade in the United States.[15] Because of increasing scrutiny over health care costs in the setting of limited resources, payors have pushed toward bundled-payment programs, with a focus on reducing unplanned readmission as one avenue to improve patient care and clinical outcomes while reducing utilization costs. Thirty-day readmission rates for TSA were reported at 5.9%, on par with readmission rates associated with total hip arthroplasty (2.4%-7.5%).[3,17] Preoperative prediction of future risk of readmission is one possible solution that can facilitate patient counseling, preoperative planning, and risk-adjusted payment. However, the prediction of patients at risk for readmission after TSA remains an unmet challenge due to the variety of factors that contribute to risk stratification.[21] There exists a need for predictive classifiers that can identify patients at risk for readmission to better inform preoperative management, shared decision making, and risk-adjusted reimbursement.

Regression based scoring such as the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) and American Joint Replacement Registry risk calculator have shown poor predictive performance in joint replacement surgery for readmission.[6,8,22] Such linear classifiers are unable to parse complex patterns that drive complication risk in patients. Machine learning (ML) is a method of using patient data from large national medical databases to identify patterns within the data that can be used to predict outcomes or events. Briefly, classifiers are trained by iterating through patient data with known labels that indicate the presence of an outcome of interest. As the classifiers continue to iterate through patient data, the algorithm coefficients are optimized to produce an algorithm that can help identify a given outcome of interest. Within orthopedics, ML classifiers are an attractive solution that have been used for prediction of complications and mortality after spinal fusion, total hip arthroplasty, total knee arthroplasty, and TSA. However, no current studies have investigated the ability of ML to predict unplanned readmission after TSA.[7,9,14,23] Moreover, although several different ML classifiers have emerged, few studies have investigated which classifiers are ideal for training on national orthopedic registries.

The purpose of this study was to evaluate ML classifiers in predicting unplanned readmission after TSA and to assess for classifiers where performance is minimally impacted by loss of data, as it occurs in real-world practice. The secondary purpose of this study was to compare the underlying predictive schemes used by different classifiers to discern why some classifiers are superior or inferior for use with orthopedic national registry data.

Materials and methods

Data collection

The ACS-NSQIP is a national surgical database that prospectively collects patient data from over 700 participating institutions.[16] All data are validated with strict adherence guidelines including routine audits to ensure high-quality data. Data from medical records, operative reports, and patient interviews are collected up to 30 days postoperatively by trained clinical reviewers. For each patient record, the NSQIP captures patient demographic/clinical characteristics, preoperative and intraoperative variables, and clinical outcomes.

Patient population and feature selection

Demographic predictor variables selected for inclusion for analysis included age >65, sex, race, body mass index >40, smoking status, functional status, preoperative weight loss >10%, and patient medical comorbidities including diabetes, congestive heart failure, chronic obstructive pulmonary disease, renal disease (acute renal failure or on dialysis), hypertension requiring medication, cancer, chronic steroid use within 30 days of surgery, coagulopathies, and American Society of Anesthesiologists' (ASA) scores $\geq$ 3. Wound infections or open wounds before surgery, as well as systemic inflammatory response syndrome (SIRS) or sepsis within 48 hours of surgery, were also included. Preoperative and intraoperative variables including lab values and type of anesthesia were used. Labs were defined as follows: leukocytosis (>10,000/mcL), low hematocrit (<30%), thrombocytopenia (<150,000/mcL), high international normalized ratio (>1.1), high creatinine (>1.3 mg/dL), high blood urea nitrogen (>30 mg/dL), and high bilirubin (>1.9 mg/dL). The clinical outcome of interest was unplanned readmission occurring within 30 days of the index operation. More information regarding each variable can be found in the ACS-NSQIP Participant Use Data File. Statistical analysis was conducted using SAS (version 9.3, Cary, NC, USA) with a 2-tailed α of 0.05. Bivariate analysis was performed to compare demographics, comorbidities, and procedure characteristics. Analysis of categorical features was performed using χ^2 tests, and continuous variables were analyzed using the Mann-Whitney U test.

Predictive classifier design

ML algorithms were developed using the Scikit-Learn (v3.7.2) package in Python. Five different classifiers were trained: support-vector machine (SVM), logistic regression (LR), random forest classifier (RF), adaptive boosting classifier (AB), and neural network (NN). These 5 classifiers were selected as they are commonly used in medical literature and have distinct pattern recognition methods.[13,16,26] Thirty-two selected features (based on data availability and orthopedic surgeon input regarding clinical relevance)

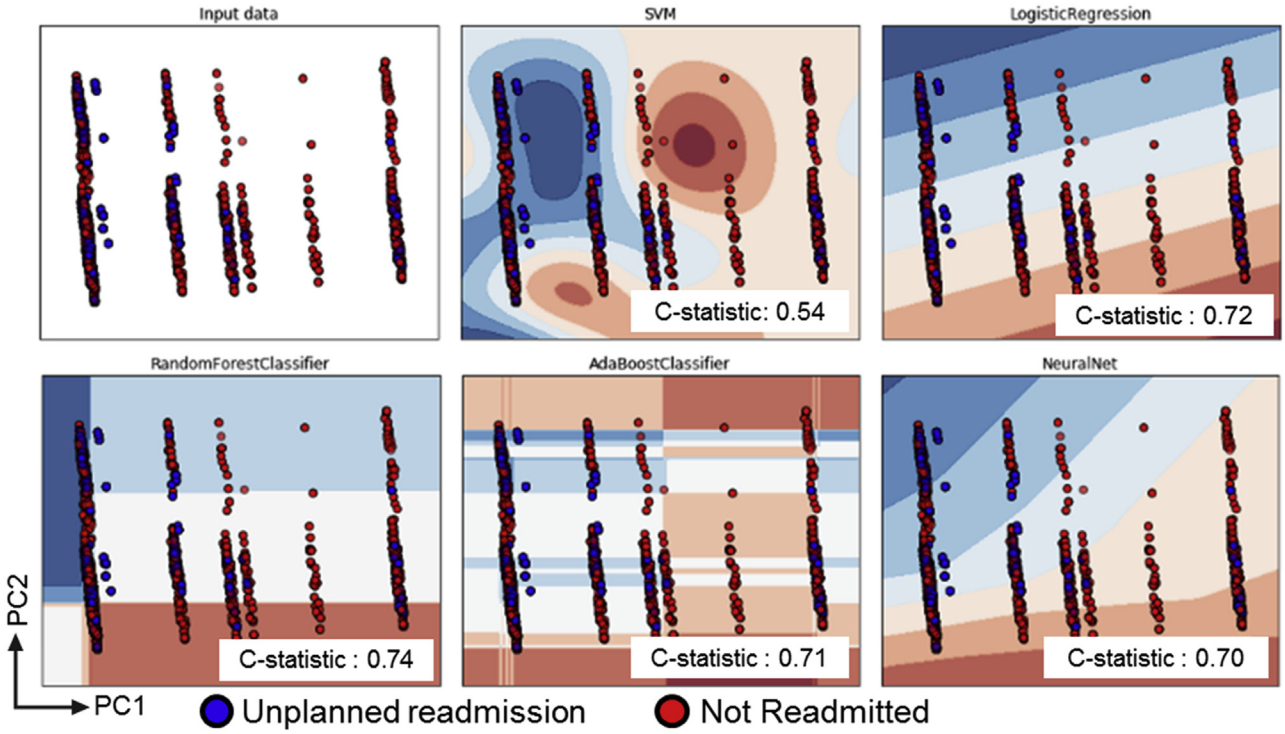


Figure 1 Classification maps of the 5 different ML classifiers. Each circle represents a patient whose position is based on demographic and clinical variables plotted using principal component analysis along axes principal component 1 (PC1) and 2 (PC2). *Blue circles* indicate patients with unplanned readmission, and *red circles* indicate patients not readmitted. Decision boundaries are displayed in the background with *blue and red* colors representing predictions for unplanned readmissions and for those not readmitted, respectively. Darker shades indicate stronger predictions, with *white* indicating an indeterminate prediction. Patients located within a given decision boundary for a classifier are predicted to have an unplanned readmission (*blue*), to not be readmitted (*red*), or it is indeterminate (*white*). With a perfect classifier, the color of each circle would match the color of the decision boundary.

from prior to surgery were used to predict unplanned readmission after TSA.

Before training, classifier weights (SVM, LR, RF, AB, NN) were initialized to random numbers. After training, the classifiers were then trained on derivation data emanating from TSA patients from 2011-2014 NSQIP cohorts, iterating through patients in order to optimize weights toward values that yield maximal classifier accuracy. Each of the five classifiers used diverse methods of pattern recognition and classification. To test which classifiers were best suited to capture patterns amongst readmitted and non-readmitted patients, classification maps were developed where each dot represents a patient (Fig. 1). Patients with unplanned readmissions clustered together supporting our hypothesis that features used for ML captured underlying differences among readmitted and non-readmitted patients. Classification maps display schemes for classifying patients from red to blue, with patients in red regions predicted to avoid readmission and patients in blue regions predicted to be readmitted. Within each group, the shade of the color indicates the strength of the prediction, with darker colors indicating stronger predictions. LR used a sequential linear pattern recognition. SVM and NN used nonlinear prediction maps. Lastly, the RF and AB classifiers used block classification maps, typical of decision-based methods.

After training on derivation data, classifiers were provided blinded data from TSA patients from the NSQIP 2015 validation cohort and asked to predict which patients would be readmitted based on input features. This step was performed to simulate real-world use.

Predictive classifier performance assessment

The clinical outcome of interest was unplanned readmission occurring within 30 days of the index operation. Primary outcomes used to evaluate classifier performance were c-statistic and f1-score, 2 commonly used metrics used to evaluate classifiers.[29] A c-statistic of 0.5 indicates a classifier functions as good as random chance whereas a c-statistic of 1 is perfect in its predictive function. In binary classifiers, the f1-score is a composite score that considers both precision and sensitivity to measure a classifier's accuracy, with a value of 1 being perfect and 0 the worst.[29]

Using the most predictive classifier tested on the 2015 TSA cohort, we stratified patients by predicted readmission risk probabilities (binned into 7 levels from 0.2 to 0.8 predicted risk, increased in increments of 0.1). For patients at each level, we assessed the proportion of patients actually readmitted vs. not readmitted. Based on the latter, a calibration plot (Fig. 2) was created demonstrating predicted readmission frequencies correlated with observed readmission.

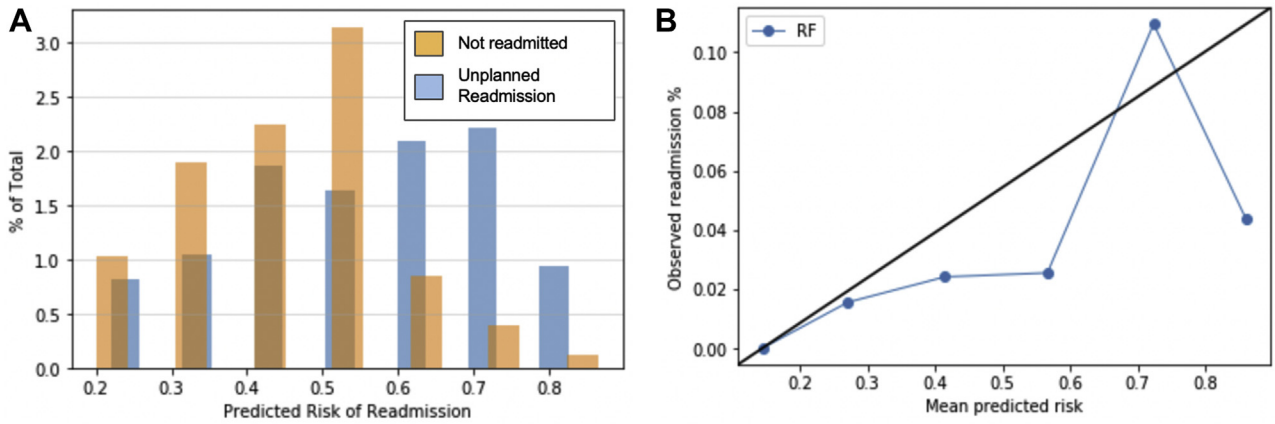


Figure 2 Calibration analysis of the random forest classifier. (**A**) Histogram of patients with unplanned readmissions and those not readmitted binned by risk of readmission predicted by the random forest classifier. (**B**) Calibration plot of mean predicted risk generated from the random forest classifier vs. observed readmission frequencies, demonstrating a correlation between predicted and observed readmission rates. *RF*, random forest.

Results

Patient characteristics, intraoperative variables, and 30-day unplanned readmission

The derivation (2011-2014) and validation (2015) TSA cohorts consisted of 5857 and 3186 patients, respectively, and were generally similar in terms of demographics, characteristics, clinical comorbidities, preoperative laboratory data, and intraoperative variables (Table I). Rates of coagulopathies (3.2% vs. 2.4%) and TSA performed for osteoarthritis etiology (69% vs. 45%) were higher in the derivation cohort, whereas rates of perioperative leukocytosis (7.5% vs. 8.7%) were lower ($P < .05$ for all). There was no statistically significant difference in 30-day unplanned readmission or mortality.

Machine learning classifier performance assessment

The RF classifier performed the best with the highest c-statistic of 0.74, positive likelihood ratio (+LR) of 1.18, negative likelihood ratio (−LR) of 0.42, and f1-score of 0.18, whereas SVM performed the worst with a c-statistic of 0.54, +LR of 1.17, −LR of 0.48, and f1-score of 0.07 (Fig. 3). In addition, the performance of SVM classifier 1 was the most sensitive to removal of a single variable, whereas the performance of RF did not dramatically decrease after the loss of a single variable (Fig. 3). Across all classifiers, the removal of the variable "elevated BUN" (blood urea nitrogen) resulted in decreased classifier performance suggesting its importance in predicting risk for 30-day unplanned readmission.

Feature importance and performance calibration plot for random forest classifier

After training of the RF classifier on the derivation cohort, weights were frozen and extracted from the RF model for analysis. Five variables achieved weights > 0.5 in descending order: high bilirubin (>1.9 mg/dL), age >65, race, chronic obstructive pulmonary disease, and ASA ≥3 (Fig. 4). Variables with the lowest contribution weights (<0.2) for RF classifier prediction included preoperative weight loss>10%, body mass index >40, diabetes, smoking, and congestive heart failure.

Risk stratification with the RF classifier identified 436 patients (13.7% of the validation cohort) with a predicted risk for readmission >0.6, which we have defined as the high-risk group. An optimal threshold of 0.6 was selected by maximizing Youden's function to identify a threshold providing optimal sensitivity and specificity.[30] Of this group, 36 (8.2%) patients were readmitted (Fig. 2). This analysis revealed that non-readmitted patients were skewed to lower predicted readmission risk, whereas readmitted patients were skewed to higher predicted readmission risk. To determine if predicted risk matched observed readmission rate, calibration analysis was performed on the RF classifier. After testing of the RF classifier on the validation cohort, predicted risk was plotted against observed readmission rate in a calibration plot. The calibration plot had a slope of 0.1 and a Pearson correlation coefficient of 0.70, supporting the predictions made by the RF classifier of match observed readmission rates after TSA (Fig. 2).

Table I Comparison of patient demographic and procedure characteristics among TSA patients in derivation and validation cohorts

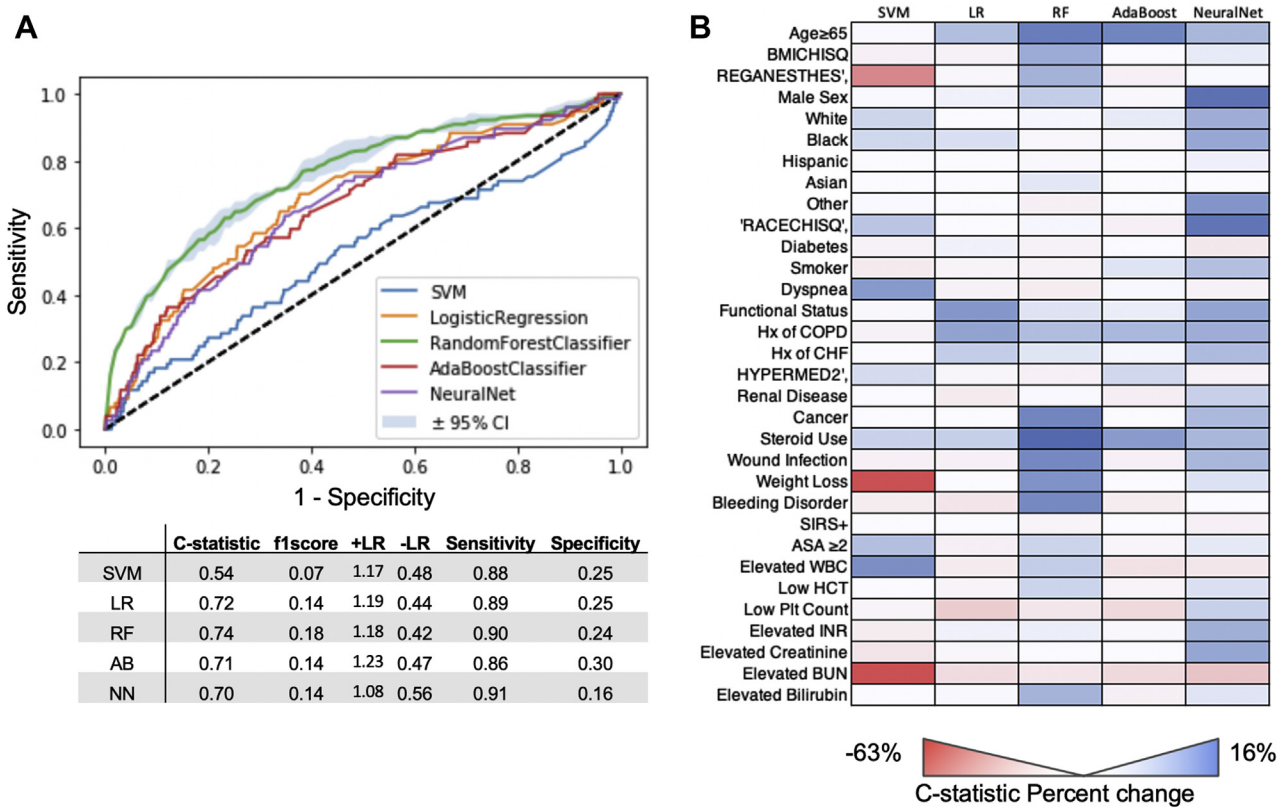
	Derivation (2011-2014), n = 5857	Validation (2015), n = 3186	*P* value
Age (mean)	69.6 (SD, 9.8)	69.1 (SD, 9.8)	.02
Male gender	2553 (43.6%)	1388 (43.6%)	.98
Dependent functional status	163 (2.8%)	87 (2.7%)	.92
BMI >40	568 (9.7%)	357 (11.2%)	.07
History of smoking	579 (9.9%)	354 (11.1%)	.83
History of diabetes	983 (16.8%)	575 (18.1%)	.13
History of pulmonary disease	382 (6.5%)	201 (6.3%)	.69
History of chronic heart failure	26 (0.44%)	15 (0.47%)	.86
Hypertension	3942 (67.3%)	2126 (66.7%)	.58
History of renal disease	31 (0.53%)	14 (0.44%)	.56
Steroids for chronic conditions	294 (5.0%)	140 (4.4%)	.18
Bleeding-causing disorders	188 (3.2%)	77 (2.4%)	.03
ASA class >2	3079 (52.6%)	1732 (54.4%)	.10
Regional anesthesia	234 (4.0%)	108 (2.4%)	.15
Operative time (mean)	115.3 (SD, 2.4)	108.9 (SD, 43.4)	<.001
Hospital LOS (mean)	2.1 (SD, 2.4)	2.0 (SD, 1.9)	<.01
Procedure etiology			
Osteoarthritis	4054 (69.2%)	1439 (45.2%)	<.0001
Traumatic arthropathy	342 (5.8%)	182 (5.7%)	.81
Inflammatory arthritis	45 (0.77%)	10 (0.31%)	.01
Other/unknown	1416 (24.2%)	1555 (48.4%)	<.0001
Compilations within 30 days			
Unplanned readmission	158 (2.7%)	91 (2.9%)	.66
Mortality	12 (0.20%)	7 (0.22%)	.88
Laboratory results within 90 d preoperatively			
Low WBC count (<4500/mcL)	301 (5.1%)	154 (4.8%)	.53
High WBC count (>10,000/mcL)	438 (7.5%)	277 (8.7%)	.04
Low hematocrit (<30%)	85 (1.5%)	52 (1.6%)	.5
Low platelets (<150,000/mcL)	298 (5.1%)	166 (5.2%)	.8
High INR (>1.1)	256 (4.4%)	146 (4.6%)	.64
Low sodium (<135 mEq/L)	355 (6.1%)	228 (7.2%)	.04
High sodium (>145 mEq/L)	58 (0.99%)	22 (0.69%)	.15
High creatinine (>1.3 mg/dL)	424 (7.2%)	221 (6.9%)	.59
High blood urea nitrogen (>30 mg/dL)	274 (4.7%)	145 (4.6%)	.78
High bilirubin (>1.9 mg/dL)	11 (0.19%)	9 (0.28%)	.36
Low albumin (<3.4 g/dL)	124 (2.1%)	85 (2.7%)	.10

TSA, total shoulder arthroplasty; *BMI*, body mass index; *ASA class*, American Society of Anesthesiologists Classification System; *LOS*, length of stay; *WBC*, white blood cell; *INR*, international normalized ratio; *SD*, standard deviation.
Data are presented as n (%) unless otherwise indicated.

Discussion

In our analysis, training of ML classifiers on 2011-2014 TSA cohorts for risk of readmission with validation against blinded data from 2015 yielded a c-statistic of 0.74 with an f1-score of 0.18, indicating acceptable risk prediction. Similar to prior studies, we found that the RF classifier outperformed other commonly used ML classifiers to predict unplanned readmission after TSA.[2,7,10,25] Furthermore, our RF classifier had good performance with a c-statistic of 0.74, which is comparable with or superior to the best performing classifiers trained to predict short-term postoperative complications after TSA developed by Gowd et al.[7]

The present study identified high bilirubin (>1.9 mg/dL), age >65, race, pulmonary disease, and ASA score≥3 as variables that contributed the greatest to RF classifier risk prediction.[11] These findings are consistent with the prior literature that shows increasing age, pulmonary disease, and ASA are predictive for the largest proportion of patients being readmitted for infection, dislocations, pneumonia, and deep vein thrombosis/pulmonary embolism.[19,28] To our knowledge, only 1 propensity score–matched study by Yin et al[32] (12,663 patients) analyzed the role of race in TSA and observed similar rates of 30-day complications and readmissions but higher mortality rates in black vs. white patients. This may be driven by psychosocial determinants of health affecting



	C-statistic	f1score	+LR	-LR	Sensitivity	Specificity
SVM	0.54	0.07	1.17	0.48	0.88	0.25
LR	0.72	0.14	1.19	0.44	0.89	0.25
RF	0.74	0.18	1.18	0.42	0.90	0.24
AB	0.71	0.14	1.23	0.47	0.86	0.30
NN	0.70	0.14	1.08	0.56	0.91	0.16

Figure 3 Random forest classifier outperforms other ML classifiers**.** (**A**) Receiver operating characteristic curves for ML classifiers evaluated on the 2015 test set with c-statistic, f1-scores, sensitivity, specificity, positive likelihood ratio (+LR), and negative likelihood ratio (−LR) displayed below. *Gray* bands indicate the 95% confidence interval for the RF classifier. (**B**) Percent change in c-statistic as a result of removal of each respective variable for each of the 5 ML classifiers. *Red* represents a negative percent change and *blue* represents a positive percent change after removal. *AB*, adaptive boosting algorithm; *ASA*, American Society of Anesthesiologists; *BUN*, blood urea nitrogen; *CHF*, congestive heart failure; *CI*, confidence interval; *COPD*, chronic obstructive pulmonary disease; *HCT*, hematocrit; *INR*, international normalized ratio; *ML*, machine learning; *NN*, neural network; *RF*, random forest; *SIRS*, systemic inflammatory response syndrome; *SVM*, support-vector machine.

patients' health literacy, access to necessary resources, and ability to comply with recommended postoperative care and follow-up. Elevated preoperative bilirubin is reflective of underlying hemolysis or hepatobiliary dysfunction, with the latter increasing potential for hemodynamic complications (eg, hypotension, hemorrhage, or hepatic ischemia).[1] Although not demonstrated in TSA, Liao et al[18] analyzed patients with compromised liver function in lumbar and hip fracture surgery and observed markedly higher rates of complications including reoperation and 30-day readmission.

Identifying patients at risk of readmission is increasingly important as health care moves toward a value-based care framework, and ML is one strategy that can be used to integrate data toward producing individualized, actionable risk-prediction scores. Recent studies have demonstrated the utility of ML to predict complications and readmission rates in high-volume spinal and joint reconstructive surgery, yet the literature is sparse with regard to TSA.[9,14,23] As the prevalence of TSA continues to grow exponentially, so does the significance of an accurate assessment of risk burden if value-based care is to be achieved. One way in which additional value can be derived is by identifying select patients as candidates for TSA at an ambulatory surgical center. Although a recent survey of 484 active American Shoulder and Elbow Surgeons members identified the greatest barrier to performing TSA in the outpatient setting was a concern for potential medical complications that may decrease procedure reimbursement, a reliable means of risk stratification via ML could help mitigate that trepidation.[4]

Furthermore, unplanned readmissions are among the costliest burdens to the health care industry, accounting for $17.4 billion among Medicare patients in 2004, and have been widely studied within the orthopedic literature as a means to curb increasing cost.[3,12,21,31] Prediction of unplanned readmission remains a complex problem, with a wide variety of factors contributing to increased risk.[21] Independent risk factors for 30-day readmission after TSA include old age (≥65), anemia, and dependent functional status; however, these risk factors likely only predict a small percentage of patients with risk of readmission.[31]

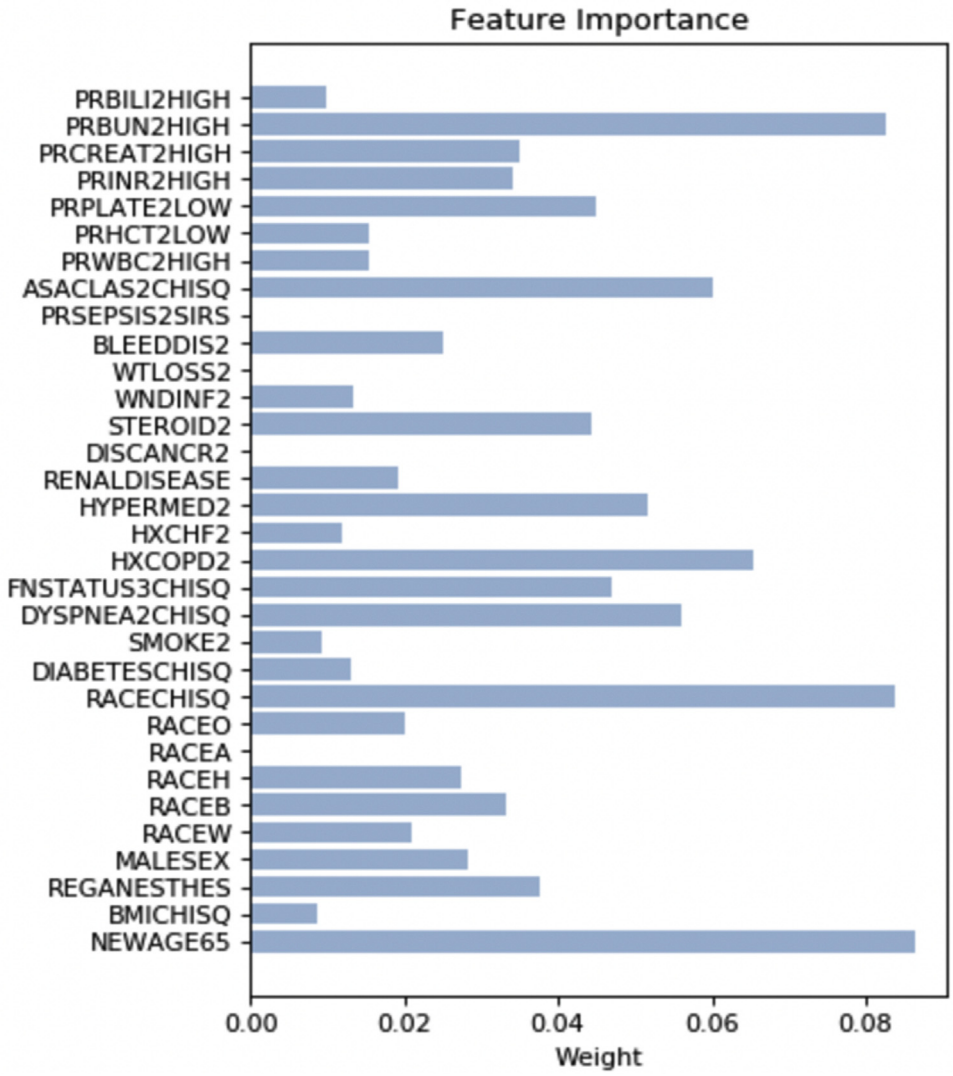


Figure 4 Random forest classifier feature weights. Feature weights assigned to variables after training of the random forest classifier on patients in the derivation cohort from 2011 to 2014. Please see the Supplementary Appendix for breakdown of the abbreviations used.

Although value-based reimbursement classifiers seek to maximize the outcomes-to-cost ratio on the population scale, this only occurs by successfully addressing individual patient needs—a reality that necessitates providers having the appropriate tools/information as well as reimbursement classifiers that adjust for individualized patient risk that cannot be modified.

Within the orthopedic literature, increasing use of ML methods to improve risk stratification of patients has highlighted a need to parse out classifiers that are best suited for orthopedic data. To investigate this, we performed a decision map analysis to characterize pattern recognition schemes of 6 commonly used classifiers in predicting unplanned readmission after TSA. The decision map analysis demonstrated that RF was best able to specifically identify patients at both an increased and decreased risk for unplanned readmission (c-statistic = 0.74). The NN and SVM classifiers created nonlinear decision maps that identified patients who had unplanned readmissions; however, both classifiers also included many patients who were not readmitted. This may be because the SVM and NN classifiers require large amounts of data to train and perform poorly when training on populations with low readmission or complication rates.[23] Logistic regression is a linear classifier, which displayed a linear decision map, with gradations in risk prediction linearly distributed from low to high, resulting in a decision map that was largely nonspecific. These analyses support the finding that RF classifiers are best suited for orthopedic datasets that comprise small training data with relatively rare adverse events or outcomes (eg, a low number of patients who experience unplanned readmissions, complications, or mortality). Indeed, within the orthopedic literature, RF classifiers routinely outperform other classifiers.[5,7,13]

In comparison to other studies, the RF classifier performed with similar performance to ML classifiers trained

to predict complications or unplanned readmissions in other high-volume orthopedic surgical procedures. Recently, ML classifiers have been used to predict length of stay in patients undergoing total hip and knee arthroplasty. Bayesian classifiers trained on 122,334 and 141,446 patients undergoing total hip and total knee arthroplasty achieved a c-statistic of 0.87 and 0.78, respectively.[23,27] In a similar study by Gowd et al,[7] ML classifiers were trained on 13,695 patients undergoing TSA to predict postoperative complications or extended length of stay with c-statistics of best-performing classifiers ranging from 0.60 to 0.77. In this study, the RF classifier achieved a c-statistic of 0.74 after training on 5857 patients within the derivation cohort demonstrating the ability to achieve good performance comparable with other studies with a smaller dataset.

The performance of ML classifiers to predict outcomes is highly dependent on class balancing (ie, the percentage of unplanned readmission).[14] When outcome frequencies are low (<10%), classifiers regrettably train to maximize accuracy and as a result fail to capture low-frequency outcomes of interest.[14] For example, in a cohort with unplanned readmissions of 5%, a classifier that classified all patients as not having an unplanned readmission would be 95% accurate. To validate if our RF was affected by such biases, we performed calibration analyses to test if classifier predictions match observed frequencies of unplanned readmissions. In examining the distribution of risk predictions for patients with and without unplanned readmissions, we observed that patients who were not readmitted had lower predicted scores, whereas patients with unplanned readmissions had higher predicted scores. Calibration plot analysis demonstrated that mean predicted risk matches observed frequency (Pearson correlation coefficient, $r = 0.70$). In other words, in a group with a mean predicted risk of unplanned readmission of 0.80, 80% would have unplanned readmissions. Although the c-statistic of the RF model was 0.74, indicating good classification performance, f1-scores were poor across all classifiers tested. f1-score is the harmonic mean of precision and recall. In a 2-class classification problem such as in this study, precision is defined as true-positives/(true-positives + false-positives), or positive predictive value (PPV), and recall is defined as true-positives/(true-positives + false-negatives), or sensitivity. Therefore, f1-score is the mean of the sensitivity and PPV, providing a metric of how well a classifier identifies all true-positives, taking into account the predictive value of each prediction. In this study, readmission rate was low, with roughly 3% of patients experiencing unplanned readmission after TSA. Because PPV is directly dependent on prevalence, if the RF classifier was tested in a cohort with increased rates of unplanned readmission, we would expect an increase in f1-score. In a study with a similar unplanned readmission rate of 3.5%, Pauly et al[24] trained an ML algorithm to predict all-cause readmission in a French medical system and observed a c-statistic of 0.74 with an f1-score of 0.13. The present study demonstrates that ML, in particular RF classifiers, can be used to reliably predict unplanned readmission from individual patient data. By implementing such classifiers, surgeons can perform individualized risk adjustment during preoperative evaluation for improved individualized care planning, risk-adjusted reimbursement, and shared decision making.

Some limitations of this study should be noted. Although we selected routinely measured variables and performed analyses of how our classifiers were affected by the loss of 1 variable, loss of 2 or more variables could significantly affect the performance of our classifier. Moreover, although the ACS-NSQIP database is a large national database covering surgical centers and hospitals throughout the United States, differences in data collection standards and rigor may affect classifier performance and generalizability to other datasets. Furthermore, as the ACS-NSQIP dataset is a broad surgical dataset, it does not contain granular information relevant to orthopedic surgery. For example, current procedural terminology codes were used to identify patients undergoing TSA; however, this does not discriminate between TSA and rTSA. In addition, continuous variables were dichotomized according to abnormal cutoffs as training on continuous data led to decreased model performance across all classifiers. Loss in performance when using continuous variables may be a result of inability of the classifiers to learn meaningful cutoffs for each feature based on the relatively small dataset. Future studies using larger datasets with continuous data that also include more granular surgical information may provide greater improvements in classifier performance. In addition, although the main strength of the classifiers discussed in this study is individualized patient stratification, the classifiers discussed do not take into consideration individualized surgeon experience. Several studies have demonstrated that patients who undergo TSA with surgeons or at hospitals with a high volume of TSA cases have reduced complications, length of stay, and readmissions. In a study by Lyman et al,[20] 60-day readmission rates were 9.5% at low-volume hospitals, whereas only 4.6% at high-volume hospitals, emphasizing the importance of institution on rates of readmission. Future studies that also incorporate institution and surgeon data may be better able to predict unplanned readmission by taking into account the whole surgical team.

Conclusion

ML is able to predict unplanned readmission after TSA in patients from a national database. Furthermore, when tested in a blinded fashion, the RF classifier outperformed other ML classifiers, with its predictions correlating best with observed frequencies. With growing datasets, ML-based classifiers may become

common place in the hospital setting, thereby allowing surgeons to better counsel patients preoperatively, deliver better individualized outcomes perioperatively, and provide greater value from TSA on a population-based scale.

Disclaimer

The authors, their immediate families, and any research foundations with which they are affiliated have not received any financial payments or other benefits from any commercial entity related to the subject of this article.

Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jse.2020.05.013.

References

1. Abbas N, Makker J, Abbas H, Balar B. Perioperative care of patients with liver cirrhosis: a review. Health Serv Insights 2017;10:1178632917691270. https://doi.org/10.1177/1178632917691270
2. Alickovic E, Subasi A. Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier. J Med Syst 2016;40:108. https://doi.org/10.1007/s10916-016-0467-8
3. Bosco JA, Karkenny AJ, Hutzler LH, Slover JD, Iorio R. Cost burden of 30-day readmissions following medicare total hip and knee arthroplasty. J Arthroplasty 2014;29:903-5. https://doi.org/10.1016/j.arth.2013.11.006
4. Brolin TJ, Cox RM, Zmistowski BM, Namdari S, Williams GR, Abboud JA. Surgeons' experience and perceived barriers with outpatient shoulder arthroplasty. J Shoulder Elbow Surg 2018;27(Suppl):S82-7. https://doi.org/10.1016/j.jse.2018.01.018
5. Durand WM, DePasse JM, Daniels AH. Predictive modeling for blood transfusion after adult spinal deformity surgery: a tree-based machine learning approach. Spine 2018;43:1058. https://doi.org/10.1097/BRS.0000000000002515
6. Edelstein AI, Kwasny MJ, Suleiman LI, Khakhkhar RH, Moore MA, Beal MD, et al. Can the American College of Surgeons risk calculator predict 30-day complications after knee and hip arthroplasty? J Arthroplasty 2015;30(Suppl):5-10. https://doi.org/10.1016/j.arth.2015.01.057
7. 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
8. Harris AHS, Kuo AC, Bozic KJ, Lau E, Bowe T, Gupta S, et al. American Joint Replacement Registry risk calculator does not predict 90-day mortality in veterans undergoing total joint replacement. Clin Orthop Relat Res 2018;476:1869. https://doi.org/10.1097/CORR.0000000000000377
9. Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori NJ. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? Clin Orthop RelatRes 2019;477:452-60. https://doi.org/10.1097/CORR.0000000000000601
10. Hsieh CH, Lu RH, Lee NH, Chiu WT, Hsu MH, Li YC. Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. Surgery 2011;149:87-93. https://doi.org/10.1016/j.surg.2010.03.023
11. Ingraham AM, Richards KE, Hall BL, Ko CY. Quality improvement in surgery: the American College of Surgeons National Surgical Quality Improvement Program approach. Adv Surg 2010;44:251-67. https://doi.org/10.1016/j.yasu.2010.05.003
12. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the Medicare fee-for-service program. N Engl J Med 2009;360:1418-28. https://doi.org/10.1056/NEJMsa0803563
13. Karhade AV, Ogink P, Thio Q, Broekman M, Cha T, Gormley WB, et al. Development of machine learning algorithms for prediction of discharge disposition after elective inpatient surgery for lumbar degenerative disc disorders. Neurosurg Focus 2018;45:E6. https://doi.org/10.3171/2018.8.FOCUS18340
14. Kim JS, Merrill RK, Arvind V, Kaji D, Pasik SD, Nwachukwu CC, et al. Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion. Spine 2018;43:853-60. https://doi.org/10.1097/BRS.0000000000002442
15. Kim SH, Wise BL, Zhang Y, Szabo RM. Increasing incidence of shoulder arthroplasty in the United States. J Bone Joint Surg Am 2011;93:2249-54. https://doi.org/10.2106/JBJS.J.01994
16. Kuo C-Y, Yu L-C, Chen H-C, Chan C-L. Comparison of models for the prediction of medical costs of spinal fusion in Taiwan diagnosis-related groups by machine learning algorithms. Healthc Inform Res 2018;24:29-37. https://doi.org/10.4258/hir.2018.24.1.29
17. Kurtz SM, Lau EC, Ong KL, Adler EM, Kolisek FR, Manley MT. Hospital, patient, and clinical factors influence 30- and 90-day readmission after primary total hip arthroplasty. J Arthroplasty 2016;31:2130-8. https://doi.org/10.1016/j.arth.2016.03.041
18. Liao J-C, Chen W-J, Chen L-H, Niu C-C, Fu T-S, Lai P-L, et al. Complications associated with instrumented lumbar surgery in patients with liver cirrhosis: a matched cohort analysis. Spine J 2013;13:908-13. https://doi.org/10.1016/j.spinee.2013.02.028
19. Lovy AJ, Keswani A, Beck C, Dowdell JE, Parsons BO. Risk factors for and timing of adverse events after total shoulder arthroplasty. J Shoulder Elbow Surg 2017;26:1003-10. https://doi.org/10.1016/j.jse.2016.10.019
20. Lyman S, Jones EC, Bach PB, Peterson MGE, Marx RG. The association between hospital volume and total shoulder arthroplasty outcomes. Clin Orthop Relat Res 2005;432:132-7. https://doi.org/10.1097/01.blo.0000150571.51381.9a
21. Mahoney A, Bosco JA, Zuckerman JD. Readmission after shoulder arthroplasty. J Shoulder Elbow Surg 2014;23:377-81. https://doi.org/10.1016/j.jse.2013.08.007
22. Manning DW, Edelstein AI, Alvi HM. Risk prediction tools for hip and knee arthroplasty. J Am Acad Orthop Surg 2016;24:19-27. https://doi.org/10.5435/JAAOS-D-15-00072
23. Navarro SM, Wang EY, Haeberle HS, Mont MA, Krebs VE, Patterson BM, et al. Machine learning and primary total knee arthroplasty: patient forecasting for a patient-specific payment model. J Arthroplasty 2018;33:3617-23. https://doi.org/10.1016/j.arth.2018.08.028
24. Pauly V, Mendizabal H, Gentile S, Auquier P, Boyer L. Predictive risk score for unplanned 30-day rehospitalizations in the French universal health care system based on a medico-administrative database. PLoS One 2019;14:e0210714. https://doi.org/10.1371/journal.pone.0210714
25. Peng S-Y, Chuang Y-C, Kang T-W, Tseng K-H. Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination. Eur J Neurol 2010;17:945-50. https://doi.org/10.1111/j.1468-1331.2010.02955.x
26. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347-58. https://doi.org/10.1056/NEJMra1814259

27. Ramkumar PN, Navarro SM, Haeberle HS, Karnuta JM, Mont MA, Iannotti JP, et al. Development and validation of a machine learning algorithm after primary total hip arthroplasty: applications to length of stay and payment models. J Arthroplasty 2019;34:632-7. https://doi.org/10.1016/j.arth.2018.12.030
28. Schairer WW, Zhang AL, Feeley BT. Hospital readmissions after primary shoulder arthroplasty. J Shoulder Elbow Surg 2014;23:1349-55. https://doi.org/10.1016/j.jse.2013.12.004
29. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, f-score and ROC: a family of discriminant measures for performance evaluation. In: Sattar A, Kang B, editors. AI 2006: advances in artificial intelligence. Berlin, Heidelberg: Springer; 2006. p. 1015-21. https://doi.org/10.1007/11941439_114
30. Unal I. Defining an optimal cut-point value in ROC analysis: an alternative approach. Comput Math Methods Med 2017;2017:3762651. https://doi.org/10.1155/2017/3762651
31. Westermann RW, Anthony CA, Duchman KR, Pugely AJ, Gao Y, Hettrich CM. Incidence, causes and predictors of 30-day readmission after shoulder arthroplasty. Iowa Orthop J 2016;36:70-4.
32. Yin C, Sing DC, Curry EJ, Abdul-Rassoul H, Galvin JW, Eichinger JK, et al. The effect of race on early perioperative outcomes after shoulder arthroplasty: a propensity score matched analysis. Orthopedics 2019;42:95-102. https://doi.org/10.3928/01477447-20190221-01