Journal of Shoulder and Elbow Arthroplasty
Volume 6: 1–8

DOI: 10.1177/24715492221075444
journals.sagepub.com/home/sea
SAGE

Development of a Machine Learning Algorithm for Prediction of Complications and Unplanned Readmission Following Primary Anatomic Total Shoulder Replacements

Sai K Devana[1], Akash A Shah[1], Changhee Lee[2], Andrew R Jensen[1], Edward Cheung[1], Mihaela van der Schaar[2,3] and Nelson F SooHoo[1]

Abstract

Background: The demand and incidence of anatomic total shoulder arthroplasty (aTSA) procedures is projected to increase substantially over the next decade. There is a paucity of accurate risk prediction models which would be of great utility in minimizing morbidity and costs associated with major post-operative complications. Machine learning is a powerful predictive modeling tool and has become increasingly popular, especially in orthopedics. We aimed to build a ML model for prediction of major complications and readmission following primary aTSA.

Methods: A large California administrative database was retrospectively reviewed for all adults undergoing primary aTSA between 2015 to 2017. The primary outcome was any major complication or readmission following aTSA. A wide scope of standard ML benchmarks, including Logistic regression (LR), XGBoost, Gradient boosting, AdaBoost and Random Forest were employed to determine their power to predict outcomes. Additionally, important patient features to the prediction models were indentified.

Results: There were a total of 10,302 aTSAs with 598 (5.8%) having at least one major post-operative complication or readmission. XGBoost had the highest discriminative power (area under receiver operating curve AUROC of 0.689) of the 5 ML benchmarks with an area under precision recall curve AURPC of 0.207. History of implant complication, severe chronic kidney disease, teaching hospital status, coronary artery disease and male sex were the most important features for the performance of XGBoost. In addition, XGBoost identified teaching hospital status and male sex as markedly more important predictors of outcomes compared to LR models.

Conclusion: We report a well calibrated XGBoost ML algorithm for predicting major complications and 30-day readmission following aTSA. History of prior implant complication was the most important patient feature for XGBoost performance, a novel patient feature that surgeons should consider when counseling patients.

Date received: 29 July 2021; revised: 23 December 2021; accepted: 5 January 2022

Introduction

Anatomic total shoulder arthroplasty (aTSA) has been shown to reliably improve pain and range of motion in patients with severe glenohumeral osteoarthritis (OA) and a functioning rotator cuff.[1] Compared to hemiarthroplasty (HA) and reverse total shoulder arthroplasty (rTSA), aTSA has been shown to have lower surgical complication and readmission rates.[2–5] Over 40,000 primary aTSA procedures are performed in the United states each year.[6] With an aging population and advancements in implants, the annual incidence of aTSA is projected to increase by up to 50% by 2025.[6,7] However, given the significant cost and morbidity inherently associated with complications and unplanned readmissions, accurate risk stratification of patients who undergo aTSA would be of great utility.

[1]David Geffen School of Medicine UCLA, Los Angeles, CA
[2]University of California, Los Angeles, CA
[3]University of Cambridge, Cambridge, UK

Corresponding author:
Sai K Devana, 10982 Roebling Ave (APT 337), Los Angeles, CA 90024.
Email:skdevana@gmail.com

Use of machine learning (ML) methods to generate prediction models has become increasingly popular within orthopaedics due to their ability to detect complex non-linear relationships and identify novel predictive factors.[8–11] Patient records, especially billing based data sets, contain large amounts of quantitative and qualitative data which introduces too many variables for traditional regression models to perform optimally.[12] Recent studies have shown that ML algorithms outperform linear models and commonly used indices in predicting outcome such as Patient Reported Outcome Measures (PROMs), readmissions, and extended length of stay.[13–17] An accurate ML based prediction model for aTSA has the potential for improving preoperative decision making, informed consent, post-operative outcomes and help guide outcome-based performance measures and reimbursement programs.

Our primary aim was to build a ML model for prediction of major perioperative complications and readmission following primary aTSA for OA. Secondarily, we aim to compare the performance of our ML models to traditional logistic regression (LR). Lastly, we aim to compare the relative importance of clinical patient features that predict outcomes in our best performing model to the most important predictive patient features from logistic regression models. We hypothesized that we will develop a ML model that outperforms LR and identify novel patient features that are important for prediction of major complications and readmissions following primary aTSA.

Methods

Data Source

Data were obtained from California's Office of Statewide Health Planning and Development (OSHPD) database, a mandatory statewide database containing codes for up to 24 diagnoses and 20 inpatient procedures per hospitalization from all licensed nonfederal hospitals in California. The OSHPD database includes patient and hospital characteristics including age, gender, race/ethnicity, insurance type, multiple comorbidities, and hospital volume. Patients in this database are assigned unique record linkage numbers that allows patients to be tracked longitudinally for complications regardless of whether future admission are at a different hospital in the database from where the index procedure was performed.

Inclusion and Exclusion Criteria

The OSHPD database was retrospectively reviewed to select patients older than 18 from October first 2015 to December 13th 2017 who underwent elective primary aTSA using International Classification of Diseases, Tenth Revision, procedural codes (ICD-10-CPS) codes. The exclusion criteria included patients with fracture of the upper extremity/shoulder girdle coded in the principal or secondary discharge diagnosis fields of index admission; concurrent revision, resurfacing, or implanted device/prosthesis removal procedure; mechanical complications coded in the principal discharge diagnosis field; malignant neoplasm of the upper

Table 1. Baseline Cohort Demographics.

Variable	All Patients (n = 10,302)
	Median (IQR)
Age (years)	71 (12)
Hospital volume†	103 (141)
	Number (%)
Male	4727 (45.88)
Race	
White	8835 (85.76)
Black	333 (3.23)
Asian / Pacific Islander	215 (2.09)
Native American	51 (0.49)
Other	777 (7.54)
Unknown	91 (0.88)
Ethnicity	
Non-Hispanic	8900 (86.39)
Hispanic	1309 (12.71)
Unknown	93 (0.90)
Insurance	
Medicare	7433 (72.15)
Private	1831 (17.77)
Medi-Cal	393 (3.81)
Workers' compensation	518 (5.02)
Other	127 (1.23)
Medical comorbidities	
Diabetes mellitus without complications	739 (7.17)
Diabetes mellitus with complications	662 (6.43)
Coronary atherosclerosis	719 (6.98)
Morbid obesity	664 (6.45)
COPD	700 (6.79)
Chronic kidney disease, mild	682 (6.62)
Chronic kidney disease, moderate	621 (6.03)
Chronic kidney disease, severe	553 (5.37)
Chronic kidney disease requiring dialysis	549 (5.33)
Vascular disease	662 (6.43)
Other circulatory disease	623 (6.05)
Acute renal failure	650 (6.31)
Cardio-respiratory failure	603 (5.85)
Major depressive or bipolar disorder	636 (6.17)
Major fracture (except skull)	574 (5.57)
Hip fracture or dislocation	554 (5.38)
Protein-calorie malnutrition	573 (5.56)
Metastatic cancer or leukemia	544 (5.28)
Complications of implants	708 (6.87)
History of prior complications	595 (5.78)
Osteoarthritis of hip or knee	737 (7.15)
Osteoporosis	667 (6.47)
History of bone/joint/muscle infection	589 (5.72)
	Mean (SD)
Number of comorbidities	0.34 (1.19)

*I*QR = interquartile range; COPD = chronic obstructive pulmonary disease; SD = standard deviation

† Cases of primary aTSA performed between 10/1/2015 and 12/13/2017

Table 2. Major Complications and Readmission.

Complications	All Patients (n = 10,302)
	Number (%)
At least one complication or readmission	598 (5.8)
Readmission within 30 days	400 (3.88)
Wound infection	157 (1.52)
Sepsis	38 (0.37)
Mechanical complication	4 (0.04)
Pneumonia	83 (0.81)
Pulmonary embolism	27 (0.26)
Surgical site bleeding	34 (0.33)
Acute myocardial infarction	23 (0.22)

extremities/shoulder girdle, bone/bone marrow or a disseminated malignant neoplasm coded in the principal discharge diagnosis fields. All principal inclusion codes were: 0RRJ0JZ, 0RRJ0KZ, 0RRK0JZ and ORRKOKZ. The extensive exclusion ICD-10 codes can be made available upon request to the authors.

Outcome and Explanatory Variables

The primary outcome of interest was any major complication or unplanned readmission after index primary aTSA (Table 2). Major complications were identified using ICD-10 codes adapted from performance measures developed by the Centers for Medicare and Medicaid (CMS) for total joint replacement.[18] These include acute myocardial infarction, pneumonia, sepsis, pulmonary embolism, surgical site bleeding, wound infection and mechanical complication. Myocardial infarction, pneumonia, and sepsis were included if the complication occurred during the index admission or within seven days of start of index admission. Pulmonary embolism was included if it occurred during the index admission or within 30 days of admission. Surgical site bleeding, wound infection and mechanical complication were included during the index admission or within 90 days. Readmission for any cause within 30 days of index aTSA was also included as an outcome.

The patient features (explanatory variables) included in or derived from the OSHPD database include patient demographic characteristics (ie age, sex, race, ethnicity, body mass index, insurance type), hospital type (community vs. teaching hospital) and patient medical comorbidities using the CMS condition categories as defined by the CMS Hierarchical Condition Category (HCC) risk adjustment model (ie diabetes, coronary artery disease, chronic obstructive pulmonary disease, malignancy, renal failure).

Model Development

We utilized 5 standard ML benchmarks that cover different classes of ML modeling as follows: LR (linear classifier), random forest[19] (a tree-based ensemble classifier), AdaBoost,[20] gradient boosting machines[21] (Gradient Boosting), and XGBoost[22] (boosting ensemble classifiers). We implemented LR, Random Forest, AdaBoost, and Gradient Boosting machines using the *scikit-learn* Python library[23] and XGBoost using the *xgboost* Python library.[22] The hyperparameters (which define the mathematical limits of an ML algorithm) of each model were selected via grid search: for LR, the coefficient for L2 regularization was chosen from a set of values in a logarithmic scale between 1e-3 to 1e3; for Random Forest, Adaboost, Gradient Boosting, and XGBoost the number of trees and the maximum depth of each tree were selected from {50, 100, 200, 300} and {2, 3, 4, 5}, respectively.

Model Evaluation

In statistical modeling, discrimination refers to how well a model distinguishes patients who developed post-operative complications and those who did not., while calibration refers to the level of agreement between prediction and the observed outcomes. We evaluated the discriminative and calibration performances of the prognostic models via 5-fold stratified cross-validation to avoid overfitting. In every cross-validation fold, the training cohort (80% of the study population) was used to derive our 5 ML benchmark models, and then a held-out testing cohort (20% of the study population) was used for performance evaluation.

Discrimination was assessed using area under the receiver operating characteristic curve (AUROC). AUROC represents the probability that a randomly selected patient who experienced an outcome was assigned a higher risk by the model than a patient who did not experience the outcome. An AUROC of 0.5 indicates that a prognostic model has no discriminative power while an AUROC of 1 indicates that a prognostic model provides perfect discrimination.[24] Calibration was assessed using Brier scores: a measure of the agreement between the observed binary outcome and the predicted probability of that outcome, which is equivalent to the mean squared error. Lower Brier scores indicates better calibration of the prognostic model.

In addition to AUROC, we also determined area under the precision-recall curve (AUPRC) values which is a useful performance metric when analyzing an imbalanced dataset; that is, a dataset where negative cases far outnumber positive cases. The precision-recall (PR) curve is constructed by plotting positive predictive value (precision) versus the sensitivity (recall). The PR curve focuses on identifying the ability of the model to correctly identify positive cases; it ignores true negatives, which is the dominant group in an imbalanced dataset.[25,26] Unlike the AUROC, the baseline AUPRC is the proportion of true positive cases. An ideal classifier predicts every positive case (perfect recall) without marking any negative case as positive (perfect precision) and will return an AUPRC of 1. Random prediction will result in the baseline

Table 3. Discrimination and Calibration.

Model	AUROC	AUPRC	Brier score
XGBoost	**0.689 ± 0.026**	0.207 ± 0.0.44	**0.051 ± 0.002**
Logistic Regression	0.662 ± 0.026	0.137 ± 0.024	0.055 ± 0
Gradient Boosting	0.687 ± 0.027	**0.214 ± 0.049**	**0.051 ± 0.002**
AdaBoost	0.677 ± 0.013	0.199 ± 0.049	0.245 ± 0.002
Random Forest	0.624 ± 0.022	0.121 ± 0.016	0.061 ± 0.001

AUPRC. The further the AUPRC is from the random prediction value, the better the model handles positive cases.

AUROC, AUPRC and Brier scores were reported as mean values with standard deviations (SD).

Feature Importance

We utilized the partial dependence function introduced by Friedman *et al.* 2001[21] to measure the importance of an individual feature by assessing the average effect in predicted risks when its value is perturbed (Appendix I). The continuous variables were standardized to zero mean and unit variance, and the categorical variables were one-hot encoded.

Results

Demographic Characteristics

Between 10/01/2015 to 12/13/2017, there was a total of 10, 302 primary aTSAs, the majority of which were females (54%). Patient age ranged from 45 to 98 years old with a median age of 73. Overall demographics and some of the most common medical comorbidities are summarized in Table 1. A total of 598 (5.8%) patients had at least one complication or readmission. There were 400 (3.9%) patients who required readmission within 30 days. The most common complications were wound infection, pneumonia and sepsis (Table 2).

Model Performance and Calibration

XGBoost demonstrated higher discrimination compared to LR (AUROC 0.689 vs. 0.662) as well as outperforming the other three standard benchmark models (Table 3). XGBoost is well-calibrated with Brier score of 0.051. The LR and standard ML models are similarly well-calibrated with the exception of AdaBoost. XGBoost and Gradient Boosting models had the highest AUPRC values of 0.207 and 0.214. These values are compared against a random classifier for this cohort of 0.058. The receiver operating characteristic curves and precision recall curves of the XGBoost and logistic regression models are depicted in Figures 1 and 2 respectively.

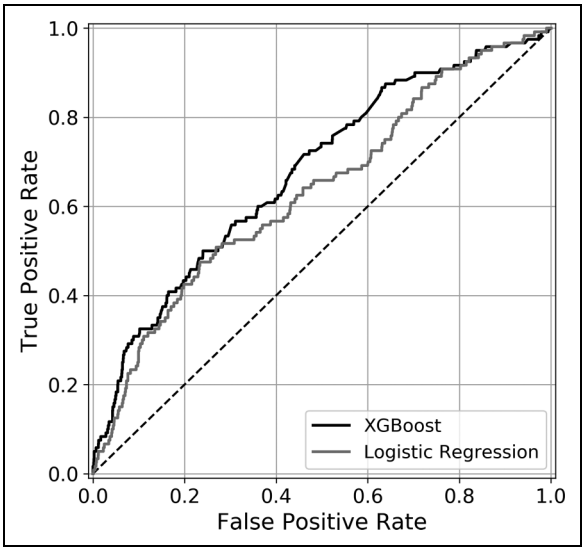

Figure 1. Area Under Receiver Operating Curve. Receiver operating characteristic curves for XGBoost and logistic regression

Relative Feature Importance

Given XGBoost had the highest discriminative performance based on AUROC, the relative importance of the categorical, continuous and top ten binary variables to the performance of the XGBoost model are listed in Table 4. A history of implant complications was the most important predictive feature for both XGBoost and LR. Specific to XGBoost, history of implant complication was more than twice as important for the model as the second most influential feature (chronic kidney disease, severe). Overall, the majority of variables that are important for XGBoost differed from those that were important for LR.

Discussion

Due to an aging population and improved outcomes, the demand for shoulder arthroplasty (including aTSA) is projected to increase drastically over the next 5 to 10 years.[6,27] With this increase in prevalence, mitigating the morbidity and costs associated with post-operative complications and readmissions is increasingly important. The purpose of this study was to create an ML algorithm to predict perioperative complications following aTSA using a statewide retrospective database. We found that XGBoost produced the most accurate predictive model and, by analyzing relative feature importance, that a history of prior implant complications was the most important predictive patient feature.

Until recently, multivariate LR has been the most prevalent modeling method used for outcome prediction.[28–30] However, ML, a subset of artificial intelligence, has grown in popularity due to advanced detection of complex nonlinear relationships and factor-factor interactions within a

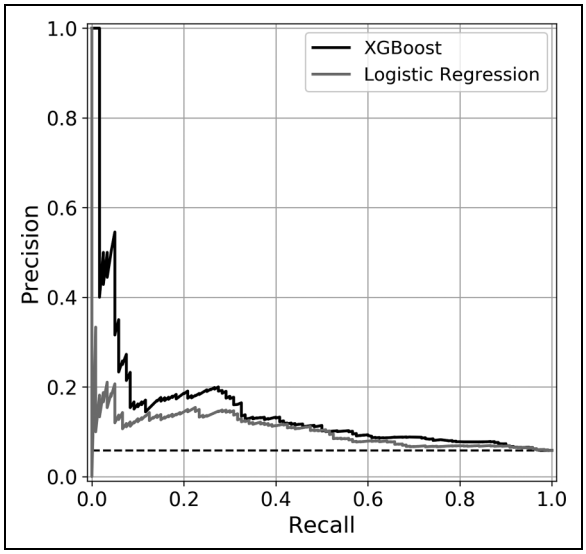

Figure 2. Area Under Precision Recall Curve. Precision Recall Curves of XGBoost and logistic regression.

given dataset.[31,32] Accordingly, ML models have been shown to outperform LR in many cases across different medical and surgical specialties.[33–35] Substantial effort has gone into utilizing ML to develop various prediction models of outcomes in orthopaedic surgery.[36–38] Clinically, developing prediction models using ML can be very useful for physicians and patients when it comes to setting expectations, preparing for particular outcomes and managing complications. Accurate ML models can also be used for allocating physician reimbursement by differentiating the high risk, complicated patients for whom increased care and costs are expected for a particular procedure.

ML has been sparingly used to predict outcomes following aTSA. Using the American College of Surgeons-National Surgical Quality Improvement Program (ACS-NASQIP) database, Gowd et. al developed an ML tool to aid in patient selection for outpatient TSA based on medical comorbidities and demographic factors.[39] With a cohort of 4500 patients, their random forest ML model was used to predict which patients had a length of stay of 1 day or less with an AUROC of 0.77. Most recently Arvind and colleagues analyzed 9043 patients who underwent primary total shoulder arthroplasty to predict 3-day unplanned readmission rates. They reported C-Statistic scores (analogous to AUROC) of 0.74 using Random Forest and 0.54 using support vector machine.

To our knowledge, this is the first study to utilize XGBoost, a gradient boosting machine learning method, in predicting complications/readmission following TSA. The model had an AUROC of 0.689, which is comparable to the aforementioned similar studies, is well-calibrated, and demonstrates superior discrimination than LR. As previously stated, AUPRC can be a useful performance metric to evaluate models built on imbalanced data set.[25,26] Our XGBoost model demonstrated good performance with an AUPRC of 0.207 compared to a random classifier of 0.058 for this cohort.

We also evaluated the relative importance of 64 different binary patient variables along with continuous variables (patient age and hospital volume) and categorical variables (patient insurance and race/ethnicity). Some ML methods may allow for detection of indirect nonlinear relationships and multivariate effects that others are not able to identify. Therefore, it is important to recognize that prediction models should not be interpreted as explanatory models. Specifically, the magnitude of feature importance should not be taken to imply causal relationships or lack thereof. Rather, inclusion of these features increases the predictive performance of the model. The most important binary feature for both the XGBoost and LR models was a history

Table 4. Relative Feature Importance for Complications or Readmission Following Primary aTSA.

Feature	Rank in XGBoost (Rank in Logistic Regression)	Change to Risk Prediction
Binary features		
Complication of implants	1 (1)	0.0297
Chronic kidney disease, severe	2 (23)	0.0117
Teaching Hospital	3 (63)	0.0111
Coronary atherosclerosis	4 (2)	0.0076
Male sex	5 (64)	0.0070
Dimentia without complications	6 (21)	0.0060
Other circulatory diseases	7 (14)	0.0059
Osteoarthritis of hip or knee	8 (31)	−0.0058
Osteoporosis	9 (3)	0.0051
Morbid Obesity	10 (18)	−0.0029
Continuous features		
Number of medical comorbidities	1 (1)	0.0650
Age	2 (3)	0.0278
Hospital volume	3 (2)	−0.0101
Insurance status		
Medicare	Reference	0
Private	1 (1)	−0098
Medical	2 (2)	0.0051
Workers Comp	3 (3)	−0.0008
Other	3 (4)	−0.0008
Race		
White	Reference	0
Asian/Pacific Islander	1 (2)	−0.0312
Black	2 (1)	0.0124
Other	3 (4)	0.0002
Native American	4 (3)	5.02E-05
Unknown	4 (5)	5.02E-05

of implant-related complication. This suggests that surgeons should weigh perioperative complications following previous orthopedic implant surgeries more significantly when counseling prior to aTSA. Severe chronic kidney disease (CKD) was also found to be the second most important predictor in our model. Dacombe *et al.* reported CKD as a predictor for increased length of stay in their multiple linear regression analysis of 640 shoulder arthroplasty cases.[40] Teaching hospital was the third most important feature, which may be secondary to more complex patients/cases being performed at teaching hospitals. Interestingly, teaching hospital as a feature was markedly less important for the LR model. Similarly, male sex and osteoarthritis of the hip or knee also were found to be important for XGBoost performance but much less so for LR. In regard to continuous variables, a patient's total number of CMS condition categories was the most important predictor. This is consistent with overall comorbidity burden contributing to the risk of complication or readmission following aTSA.

That there were significant differences in feature importance between XGBoost and LR underlines the fact that advanced machine learning methods treat the same features very differently than traditional LR-based methods. For example, male sex was the fifth most influential variable for our XGBoost algorithm whereas it was the 64th most influential variable for LR. This is possibly due to the ability of ML methods to capture occult relationships between variables that LR is unable to detect.

The retrospective nature of this study inherently lends itself to limitations. Though a de-identified state-wide code-based database has a large patient sample, the patient features and outcomes collected are limited. The reliance on ICD diagnoses and procedure codes is less reliable than thorough chart review. Code based searches of databases are dependent on accurate coding and can lead to exclusion of a patient of interest or underestimation of outcomes. With this database we were unable to assess mortality, patient-reported functional outcomes and patient satisfaction. Due to the low complication rate found in this cohort, our data may be imbalanced. However, we believe that predictive models trained with an artificially balanced dataset cannot be directly used in a clinical setting as they will be inherently poorly calibrated. To better address the concern of imbalanced data, we evaluated the five prognostic models in terms of area under precision-recall curve. Along the same lines, due to the low overall complication rate, secondary analysis of individual complications/outcomes was beyond the scope of our model (though we certainly recognize the clinical importance of such results). Lastly, we must acknowledge the black box nature of ML algorithms that can lead to non-physiologic predictors or predictors that effect a very small portion of the cohort having high significance which highlights that ML provides predictive modeling at the expense of statistical inference.

Conclusion

Here we show that the use of ML modeling, specifically using XGBoost allows for prediction of major complications and readmission following aTSA from state-wide claims-based retrospective data. Our model is well calibrated and superior in performance to a traditional LR model. Based on our relative feature importance, shoulder arthroplasty surgeons should inquire and consider a history of prior implant related complications during pre-operative counseling of patients. While further studies are needed to externally validate this model, we hope that this tool can be a building block for physicians to identify modifiable risk factors and help with preoperative counseling, managing patient expectations, informed consent, shared decision making and potentially be useful for risk-adjustment in reimbursement programs.

Acknowledgements

All the authors have no financial relationships or conflict of interests to disclose. This work was exempt from IRB approval

Author's Note

Changhee Lee, Department of Artificial Intelligence, Chung-Ang University.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was supported in part by the H. H. Lee grant through the UCLA department of Orthopedics. [HHLEE FAU 441489 to 3H-62252].

Ethical Approval

Not applicable, because this article does not contain any studies with human or animal subjects.

Informed Consent

Not applicable, because this article does not contain any studies with human or animal subjects.

Trial Registration

Not applicable, because this article does not contain any clinical trials.

ORCID iDs

Sai K Devana https://orcid.org/0000-0002-7264-2307
Changhee Lee https://orcid.org/0000-0002-8681-4739

Supplemental material

Supplemental material for this article is available online.

References

1. Simovitch RW, Friedman RJ, Cheung E V, et al. Rate of improvement in clinical outcomes with anatomic and reverse total shoulder arthroplasty. *J. Bone Jt. Surg. [Internet]*. 2017 Nov 1;99(21):1801–1811. Available from: https://journals.lww.com/00004623-201711010-00003. doi:10.2106/JBJS.16.01387.
2. Chin PYK, Sperling JW, Cofield RH, Schleck C. Complications of total shoulder arthroplasty: are they fewer or different? *J. Shoulder Elb. Surg. [Internet]*. 2006 Jan;15(1):19–22. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1058274605001564. doi:10.1016/j.jse.2005.05.005.
3. Floyd SB, chapman CG, thigpen CA, brooks JM, hawkins RJ, tokish JM. Shoulder arthroplasty in the US medicare population: a 1-year evaluation of surgical complications, hospital admissions, and revision surgery. *JSES Open Access [Internet]*. 2018 Mar;2(1):40–47. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2468602617300542. doi:10.1016/j.jses.2017.10.002.
4. Jiang JJ, Toor AS, Shi LL, Koh JL. Analysis of perioperative complications in patients after total shoulder arthroplasty and reverse total shoulder arthroplasty. *J. Shoulder Elb. Surg. [Internet]*. 2014 Dec;23(12):1852–1859. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1058274614002298. doi:10.1016/j.jse.2014.04.008.
5. Mahoney A, Bosco JA, Zuckerman JD. Readmission after shoulder arthroplasty. *J. Shoulder Elb. Surg. [Internet]*. 2014 Mar;23(3):377–381. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1058274613004345. doi:10.1016/j.jse.2013.08.007.
6. Wagner ER, Farley KX, Higgins I, Wilson JM, Daly CA, Gottschalk MB. The incidence of shoulder arthroplasty: rise and future projections compared with hip and knee arthroplasty. *J. Shoulder Elb. Surg. [Internet]*. 2020 Dec;29(12):2601–2609. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1058274620303517. doi:10.1016/j.jse.2020.03.049.
7. Khatib O, Onyekwelu I, Yu S, Zuckerman JD. Shoulder arthroplasty in New York state, 1991 to 2010: changing patterns of utilization. *J. Shoulder Elb. Surg*. 2015;24(10):e286–e291. doi:10.1016/j.jse.2015.05.038
8. Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. *Front. Bioeng. Biotechnol. [Internet]*. 2018 Jun 27;6, Available from: https://www.frontiersin.org/article/10.3389/fbioe.2018.00075.
9. Karhade AV, Thio QCBS, Ogink PT, et al. Development of machine learning algorithms for prediction of 30-Day mortality after surgery for spinal metastasis. *Neurosurgery [Internet]*. 2019 Jul 1;85(1):E83–E91. Available from: https://academic.oup.com/neurosurgery/article/85/1/E83/5200909. doi:10.1093/neuros/nyy469.
10. Shah AA, Devana SK, Lee C, Kianian R, van der Schaar M, SooHoo NF. Development of a novel, potentially universal machine learning algorithm for prediction of complications after total Hip arthroplasty. *J. Arthroplasty [Internet]*. 2020 Dec. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0883540320313000. doi:10.1016/j.arth.2020.12.040
11. Shah AA, Karhade A V, Bono CM, Harris MB, Nelson SB, Schwab JH. Development of a machine learning algorithm for prediction of failure of nonoperative management in spinal epidural abscess. *Spine J. [Internet]*. 2019 Oct;19(10):1657–1665. Available from: https://linkinghub.elsevier.com/retrieve/pii/S152994301930169X. doi:10.1016/j.spinee.2019.04.022.
12. Murdoch TB, Detsky AS. The inevitable application of Big data to health care. *JAMA [Internet]*. 2013 Apr 3;309(13):1351. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2013.393. doi:10.1001/jama.2013.393.
13. Arvind V, London DA, Cirino C, Keswani A, Cagle PJ. Comparison of machine learning techniques to predict unplanned readmission following total shoulder arthroplasty. *J. Shoulder Elb. Surg. [Internet]*. 2021 Feb;30(2):e50–e59. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1058274620304468. doi:10.1016/j.jse.2020.05.013.
14. Ehlers AP, Roy SB, Khor S, et al. Improved risk prediction following surgery using machine learning algorithms. *eGEMs (Generating Evid. Methods to Improv. patient outcomes)*. 2017;5(2):3. doi:10.13063/2327-9214.1278
15. Gowd AK, Agarwalla A, Amin NH, et al. Construct validation of machine learning in the prediction of short-term postoperative complications following total shoulder arthroplasty. *J. Shoulder Elb. Surg. [Internet]*. 2019 Dec;28(12):e410–e421. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1058274619303520. doi:10.1016/j.jse.2019.05.017.
16. Huber M, Kurz C, Leidl R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. *BMC Med. Inform. Decis. Mak*. 2019;19(1):3. doi:10.1186/s12911-018-0731-6
17. Hyer JM, White S, Cloyd J, et al. Can We improve prediction of adverse surgical outcomes? Development of a surgical complexity score using a novel machine learning technique. *J. Am. Coll. Surg*. 2020;230(1):43–52. doi:10.1016/j.jamcollsurg.2019.09.015
18. (YNHHSC/CORE) YNHHSC-C for OR and E. 2020 Procedure-Specific Complication Measure Updates and Specifications Report: Elective Primary Total Hip Arthroplasty (THA) and/or Total knee Arthroplasty (TKA) - Version 9.0. 2020.
19. Breiman L. Random forests. *Mach. Learn*. 2001;45(6):5–32. doi:10.1017/CBO9781107415324.004
20. Rätsch G, Onoda T, Müller KR. Soft margins for AdaBoost. *Mach. Learn*. 2001;42(3):287–320. doi:10.1023/A:1007618119488
21. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann. Stat*. 2001;29(5):1189–1232.
22. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. doi:10.1145/2939672.2939785.
23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: machine learning in python. *J. Mach. Learn. Res*. 2011;12(1):2825–2830. doi:10.1145/2786984.2786995
24. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med*. 2003;29(7):1043–1051. doi:10.1007/s00134-003-1761-8
25. Ozenne B, Subtil F, Maucort-Boulch D. The precision-recall curve overcame the optimism of the receiver operaitng characteristic curve in rare diseases. *J. Clin. Epidemiol*. 2015;68(8):855–859.

26. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
27. Padegimas EM, Maltenfort M, Lazarus MD, Ramsey ML, Williams GR, Namdari S. Future patient demand for shoulder arthroplasty by younger patients: national projections. *Clin. Orthop. Relat. Res. [Internet]*. 2015 Jun;473(6):1860–1867. Available from: https://journals.lww.com/00003086-201506000-00003. doi:10.1007/s11999-015-4231-z.
28. Bohl DD, Shen MR, Kayupov E, Della Valle CJ. Hypoalbuminemia independently predicts surgical site infection, pneumonia, length of stay, and readmission after total joint arthroplasty. *J. Arthroplasty*. 2016;31(1):15–21. doi:10.1016/j.arth.2015.08.028
29. Bozic KJ, Ong K, Lau E, et al. Estimating risk in medicare patients with THA: an electronic risk calculator for periprosthetic joint infection and mortality hip. *Clin. Orthop. Relat. Res*. 2013;471(2):574–583. doi:10.1007/s11999-012-2605-z
30. Schoenfeld AJ, Carey PA, Cleveland AW, Bader JO, Bono CM. Patient factors, comorbidities, and surgical characteristics that increase mortality and complication risk after spinal arthrodesis: a prognostic study based on 5,887 patients. *Spine J. [Internet]*. 2013 Oct;13(10):1171–1179. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1529943013002714. doi:10.1016/j.spinee.2013.02.071.
31. Chen JH, Asch SM. Machine learning and prediction in medicine — beyond the peak of inflated expectations. *N. Engl. J. Med. [Internet]*. 2017 Jun 29;376(26):2507–2509. Available from: http://www.nejm.org/doi/10.1056/NEJMp1702071. doi:10.1056/NEJMp1702071.
32. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann. Surg. [Internet]*. 2018 Jul;268(1):70–76. Available from: https://journals.lww.com/00000658-201807000-00013. doi:10.1097/SLA.0000000000002693.
33. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature [Internet]*. 2017 Feb 2;542(7639):115–118. Available from: http://www.nature.com/articles/nature21056. doi:10.1038/nature21056.
34. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal Fundus photographs. *JAMA [Internet]*. 2016 Dec 13;316(22):2402. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2016.17216. doi:10.1001/jama.2016.17216.
35. Menden MP, Iorio F, Garnett M, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One [Internet]*. 2013 Apr 30;8(4):e61318. Available from: https://dx.plos.org/10.1371/journal.pone.0061318. doi:10.1371/journal.pone.0061318.
36. Edelstein AI, Kwasny MJ, Suleiman LI, et al. Can the American college of surgeons risk calculator predict 30-Day complications after knee and Hip arthroplasty? *J. Arthroplasty*. 2015;30(9Supp):5–10. doi:10.1016/j.arth.2015.01.057
37. Harris AHS, Kuo AC, Bozic KJ, 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(9):1869–1875. doi:10.1097/CORR.0000000000000377
38. Romine LB, May RG, Taylor HD, Chimento GF. Accuracy and clinical utility of a peri-operative risk calculator for total knee arthroplasty. *J. Arthroplasty*. 2013;28(3):445–448. doi:10.1016/j.arth.2012.08.014
39. Biron DR, Sinha I, Kleiner JE, et al. A novel machine learning model developed to assist in patient selection for outpatient total shoulder arthroplasty. *J. Am. Acad. Orthop. Surg. [Internet]*. 2020 Jul 1;28(13):e580–e585. Available from: https://journals.lww.com/10.5435/JAAOS-D-19-00395. doi:10.5435/JAAOS-D-19-00395.
40. Dacombe P, Harries L, McCann P, et al. Predictors of length of stay following shoulder arthroplasty in a high-volume UK centre. *Ann. R. Coll. Surg. Engl. [Internet]*. 2020 Sep;102(7):493–498. Available from: https://publishing.rcseng.ac.uk/doi/10.1308/rcsann.2020.0052. doi:10.1308/rcsann.2020.0052.

Appendix I

We utilized the partial dependence function introduced in Friedman *et al.* 2001 to measure the importance of an individual feature by assessing the average effect in predicted risks when its value is perturbed. More specifically, x_c is a chosen target feature in the set of input features $\mathcal{X}$ and $\mathcal{X}_{\setminus c}$ be its complement, ie, $\mathcal{X} = \mathcal{X}_{\setminus c} \cup_c^x$, and $r(\mathcal{X}) = r(\mathcal{X}_{\setminus c}, x_c)$ be the predicted risk by our trained model. Then, we define the feature importance score for an individual feature x_c by averaging $r(\mathcal{X}_{\setminus c}, x_c = 1) - r(\mathcal{X}_{\setminus c}, x_c = 0)$ for binary features and $r(\mathcal{X}_{\setminus c}, x_c = \max(x_c)) - r(\mathcal{X}_{\setminus c}, x_c = \min(x_c))$ where $\max(x_c)$ and $\min(x_c)$ are the maximum and minimum of feature x_c for continous variables. For categorical variables: we define feature importance of category $b \in \{1, \cdots, B\}$ as $r(\mathcal{X}_{\setminus c}, x_c = b) - r(\mathcal{X}_{\setminus c}, x_c = \text{mode}(x_c))$ where $\text{mode}(x_c)$ indicates the most frequency category of feature x_c.