HIP ARTHROPLASTY

Development and internal validation of machine learning algorithms for predicting complications after primary total hip arthroplasty

Kyle N. Kunze[1,4] · Aditya V. Karhade[2] · Evan M. Polce[3] · Joseph H. Schwab[2] · Brett R. Levine[1]

Abstract

Introduction Complications after total hip arthroplasty (THA) may result in readmission or reoperation and impose a significant cost on the healthcare system. Understanding which patients are at-risk for complications can potentially allow for targeted interventions to decrease complication rates through pursuing preoperative health optimization. The purpose of the current was to develop and internally validate machine learning (ML) algorithms capable of performing patient-specific predictions of all-cause complications within two years of primary THA.

Methods This was a retrospective case–control study of clinical registry data from 616 primary THA patients from one large academic and two community hospitals. The primary outcome was all-cause complications at a minimum of 2-years after primary THA. Recursive feature elimination was applied to identify preoperative variables with the greatest predictive value. Five ML algorithms were developed on the training set using tenfold cross-validation and internally validated on the independent testing set of patients. Algorithms were assessed by discrimination, calibration, Brier score, and decision curve analysis to quantify performance.

Results The observed complication rate was 16.6%. The stochastic gradient boosting model achieved the best performance with an AUC = 0.88, calibration intercept = 0.1, calibration slope = 1.22, and Brier score = 0.09. The most important factors for predicting complications were age, drug allergies, prior hip surgery, smoking, and opioid use. Individual patient-level explanations were provided for the algorithm predictions and incorporated into an open access digital application: https://sorg-apps.shinyapps.io/tha_complication/

Conclusions The stochastic boosting gradient algorithm demonstrated good discriminatory capacity for identifying patients at high-risk of experiencing a postoperative complication and proof-of-concept for creating office-based applications from ML that can perform real-time prediction. However, this clinical utility of the current algorithm is unknown and definitions of complications broad. Further investigation on larger data sets and rigorous external validation is necessary prior to the assessment of clinical utility with respect to risk-stratification of patients undergoing primary THA.

Level of evidence III, therapeutic study.

Keywords Machine learning · Predictive modeling · Total hip arthroplasty · THA · Complications

✉ Kyle N. Kunze
kylekunze7@gmail.com

1. Department of Orthopaedic Surgery, Rush University Medical Center, Chicago, IL, USA

2. Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

3. School of Medicine and Public Health, University of Wisconsin, Madison, WI, USA

4. Department of Orthopaedic Surgery, Hospital for Special Surgery, New York, NY, USA

Introduction

The implementation of bundled payment models as an economic strategy to limit financially undesirable health care costs associated with total joint arthroplasty procedures has become commonplace among hospitals throughout the United States [1, 2]. A subsequent result of such policy has been increasing focus on the ability to predict and anticipate patient outcomes, the selective recruitment of patients, and promotion of preoperative health optimization based on risk factors. Procedure-associated complications after total hip arthroplasty (THA) is a significant potential source of increased

health-care expenditures secondary to hospital and emergency room readmissions in addition to the possible need for revision arthroplasty [2]. Therefore, understanding which patients are at-risk for complications during the preoperative period would potentially allow for targeted intervention during a time when health optimization could be performed in order to possibly decrease this complication risk.

Much of the current knowledge regarding risk factors associated with complications after THA is limited to associations identified in studies that are not hypothesis-driven but found through testing many random variables [3–8]. Some of these risk factors include age [7, 9], number of comorbidities [7, 9], body mass index, [5, 10], smoking [6, 11], sex [12], and diabetes mellitus [11]. Few studies have sought to develop risk models incorporating sets of such known preoperative risk factors in order to understand and predict which patients are at higher risk for complications in the postoperative period [13], while other prediction models have been limited by using intraoperative data and therefore do not allow for preoperative optimization [14]. As such, there is currently a large pool of potential risk factors and a limited number of prediction models utilizing solely preoperative and modifiable risk factors. Developing and cross-validating models with the smallest number of the most important factors would be of great clinical utility to hip and knee surgeons. This stratification model would benefit patients by providing them with the opportunity to optimize their health prior to undergoing THA.

The application of machine learning is a powerful statistical instrument capable of determining patient-specific factors which influence the probability of a patient experiencing a complication after primary THA. Furthermore, machine learning allows for the development of clinical decision-making tools, which can be used in office-based settings to help discuss risk stratification with patients [15–17]. This may assist orthopaedic surgeons to better determine which patients may need further optimization prior to undergoing THA. The purposes of the current study were to (1) develop and internally validate machine learning algorithms capable of predicting all-cause complications within two years of primary THA, and (2) to use these algorithms to determine which preoperative factors are important in predicting all-cause complications after primary THA. The authors hypothesized that best performing machine learning algorithm would allow for both excellent prediction and an interpretable explanation of how factors specific to individual patients influenced the model decision making.

Methods

Patient selection

Following institutional board approval, data was obtained retrospectively from the electronic medical records of patients who underwent primary total hip arthroplasty by one fellowship-trained surgeons at one large academic and two community hospitals. The timeframe for patient inclusion was between January 2014 and January 2016. Exclusion criteria included etiology of degenerative hip osteoarthritis that was inflammatory, infectious, post-traumatic, acute femoral neck fracture, or related to osteonecrosis, patients undergoing revision THA, and patients with less than two-year follow-up. Overall, 616 patients met the inclusion criteria and had a median age of 62 (interquartile range [IQR] 54–70) years. A total of 352 (57.1%) patients were female. Additional demographic and clinical outcome information is displayed in Table 1. A minimum of 100 patients has been demonstrated to be an appropriate sample size for machine learning analyses and associated predictive analytics, and therefore the current sample of 616 patients was deemed valid [18, 19].

Primary outcome

The primary outcome was all-cause complications within the two-year follow-up period. Complications were considered all events classified to be either medical or orthopaedic (Table 2). Medical complications included post-operative myocardial infarction, pulmonary embolism, deep vein thrombosis, atrial fibrillation, and anemia requiring blood transfusion. Orthopaedic complications included nerve palsy, hematoma formation, heterotopic ossification, hip squeaking, wound abscesses or dehiscence, periprosthetic infection, intra- and post-operative fractures, dislocations, leg-length discrepancy, aseptic loosening, and atraumatic, return to the emergency department or readmission for any complaint related to the operative hip, and reoperations.

Table 1 Characteristics of study population, $n = 616$

Variable	n (%) | median (IQR)
Age (years)	62 (54–70)
Female sex	352 (57.1)
Body mass index (kg/m²)	29.7 (25.5–36.3)
Preoperative opioid use	169 (27.5)
Prior or current smoking history	174 (28.3)
Diabetes	87 (14.2)
Drug allergies	125 (23.6)
Any comorbidities	549 (89.4)
Prior contralateral or ipsilateral hip procedure	112 (18.2)
Contralateral	71 (63.4%)
Ipsilateral	41 (36.6%)
Preoperative Health State	45 (40–70)
Preoperative MHHS	40 (27–52)
Postoperative complications	102 (16.6)

MHHS modified Harris Hip Score

Table 2 Surgical and medical complications

Complication	Frequency (%)
Surgical	
Wound complication	10 (1.6)
Periprosthetic joint infection	7 (1.1)
Reoperations including revision THA	7 (1.1)
Return to emergency room related to THA	6 (0.97)
Prosthetic dislocation	5 (0.81)
Hematoma formation	5 (0.81)
Nerve palsy	4 (0.65)
Periprosthetic fracture	3 (0.49)
Implant fracture or hardware failure	2 (0.32)
Heterotopic ossification	2 (0.32)
Aseptic loosening	2 (0.32)
Audible squeaking of hip joint	2 (0.32)
Mortality	2 (0.32)
Implant dissociation	1 (0.16)
Medical	
Anemia requiring blood transfusion	14 (2.3)
Hyponatremia	12 (1.9)
Deep vein thrombosis	7 (1.1)
Atrial fibrillation	5 (0.81)
Pulmonary embolism	4 (0.65)
Myocardial infarction	2 (0.32)
Total	102 (16.6)

Frequency (%) of complications reported as proportion of complications occurring in entire cohort of patients, $n = 616$

Candidate variables

Candidate variables were collected prospectively before THA and stored in a secure clinical repository. Candidate variables are listed in Table 1, with the rates of missing data as follows: preoperative opioid use ($n = 2$, 0.32%), smoking history ($n = 2$, 0.32%), diabetes at time of surgery ($n = 2$, 0.32%), drug allergies ($n = 86$, 14.0%), presence of one or more comorbidities ($n = 2$, 0.32%), preoperative health state ($n = 26$, 4.2%), preoperative modified Harris Hip Score (mHHS) ($n = 17$, 2.8%), preoperative hip flexion ($n = 197$, 32.0%). Hip flexion was the only variable with greater than 30% missing data and was consequently excluded [20, 21]. These PROMs included the patient reported health state (PRHS) [22] and the modified Harris Hip Score (mHHS) [23] Prior to analysis, missingness of data was explored and determined to be missing at random and appropriate for multiple imputation. The current analysis applied multiple imputation and predictive mean matching with the "mice" package in R (R Foundation for Statistical Computing, Vienna, Austria) [24]. Following imputation, recursive feature elimination (RFE) with random forest algorithms were used to determine the combination of variables with

the highest predictive value that optimized algorithm performance through a process of backwards elimination.

Algorithm construction and performance assessment

The machine algorithm development methodology and data analysis had been previously described in detail [25, 26]. Briefly, five novel algorithms were constructed on a training set of patients (80% of initial cohort) using three iterations of tenfold cross-validation. Standardized metrics of model performance including (1) calibration (calibration plot, intercept, slope) [27, 28], (2) decision curve analysis [29, 30], (3) Brier score [31], and (4) discrimination (area under receiver operating curve), were used to comparatively evaluate model performance of both the training and testing sets.

Exploration of patient-specific model explanations

Local interpretable model-agnostic explanations (LIME) depict the decision-making process of machine learning algorithms and were used to demonstrate how the best performing algorithm explained prediction on a patient-by-patient basis [32]. Using LIME, an open access digital web application was developed with the capacity to provide both predictions and explanations at the individual patient level [33]. This application is freely accessible: https://sorg-apps. shinyapps.io/tha_complication/. Given that external validation was not performed in the present study, the application in its current form merely constitutes an educational tool and open-access source to the developed algorithms.

The Anaconda Distribution (Anaconda, Inc., Austin, Texas), R (The R Foundation, Vienna, Austria), RStudio (RStudio, Boston, MA), and Python (Python Software Foundation, Wilmington, Delaware) were used for data analysis. Predictive modeling development and testing was performed under guidelines set forth by Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines and the Guidelines for Developing and Reporting Machine Learning Models in Biomedical Research were followed for this analysis [34, 35].

Results

Final variable selection

The combination of variables identified for algorithm development through recursive feature selection that optimized predictive performance were comorbidities, preoperative opioid use greater than three months, current smoking, prior hip surgery, drug allergies, and age (Fig. 1B).

Algorithm selection and model performance

Cross-validation of the training set ($n = 494$) demonstrated that the AUC ranged from 0.81 to 0.92, the calibration intercept ranged from −0.21 to 5.09, the calibration slope ranged from 0.8 to 3.49, and the Brier score ranged from 0.08 to 0.12 (Table 3).

In the testing set, the AUC ranged from 0.77 to 0.93, the calibration intercept ranged from −0.50 to 2.11, the calibration slope ranged from 0.89 to 1.22, and the Brier score ranged from 0.08 to 0.16 (Table 4). The algorithm with the best performance was the stochastic gradient boosting model with AUC 0.88, calibration intercept 0.103, calibration slope 1.22, and Brier score 0.09. The most important factors for prediction of complications were age, documented drug allergies, prior ipsilateral hip surgery, smoking, and preoperative opioid use (Figs. 1A and 2A). The stochastic gradient boosting model resulted in greater net benefit compared to the default strategies of changes for all patients, for no patients, or changes based on age alone as demonstrated by the decision curve analysis (Fig. 2B).

Potential application and utility of machine learning using patient specific explanations

An individual patient-specific risk explanation demonstrating the potential utility of using this risk stratification model

Fig. 1 **A** Receiver operative curve demonstrating discrimination of stochastic gradient boosting algorithm. **B** Global variable importance plot, with variables ranked in decreasing order of importance
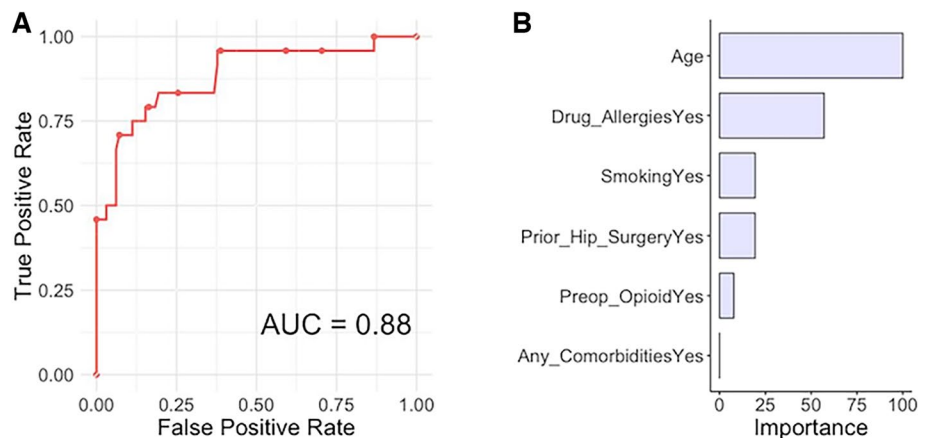
Table 3 Algorithm performance on cross-validation of training set, $n = 494$, mean (95% confidence interval)

Metric	Stochastic gradient boosting	Random forest	Support vector machine	Neural network	Elastic-net penalized logistic regression
AUC	0.86 (0.83, 0.89)	0.92 (0.89, 0.94)	0.83 (0.80, 0.86)	0.85 (0.82, 0.87)	0.81 (0.78, 0.84)
Calibration intercept	0.02 (−0.23, 0.26)	−0.21 (−0.42, 0.00)	5.09 (−3.02, 13.2)	1.05 (0.60, 1.49)	0.06 (−0.08, 0.21)
Calibration slope	1.03 (0.84, 1.22)	0.80 (0.68, 0.92)	3.49 (−0.41, 7.40)	1.10 (0.94, 1.28)	1.11 (0.96, 1.27)
Brier score	0.10 (0.09, 0.11)	0.08 (0.07, 0.09)	0.11 (0.10, 0.12)	0.12 (0.11, 0.12)	0.12 (0.11, 0.12)

Null model Brier score = 0.16

AUC area under the curve

Table 4 Algorithm performance in independent testing set (95% confidence interval), $n = 122$

Metric	Stochastic gradient boosting	Random forest	Support vector machine	Neural network	Elastic-net penalized logistic regression
AUC	0.88 (0.77, 0.95)	0.91 (0.67, 0.96)	0.77 (0.64, 0.87)	0.89 (0.82, 0.94)	0.93 (0.87, 0.96)
Calibration intercept	0.103 (−0.45, 0.65)	−0.50 (−1.27, 0.27)	0.02 (−0.51, 0.51)	2.11 (1.60, 2.63)	0.19 (−0.32, 0.69)
Calibration slope	1.22 (0.75, 1.69)	0.89 (0.41, 1.37)	0.92 (0.48, 1.37)	1.59 (0.91, 2.27)	1.65 (1.03, 2.26)
Brier score	0.09 (0.06, 0.13)	0.08 (0.05, 0.13)	0.13 (0.09, 0.18)	0.16 (0.11, 0.22)	0.11 (0.07, 0.14)

Null model Brier score = 0.16
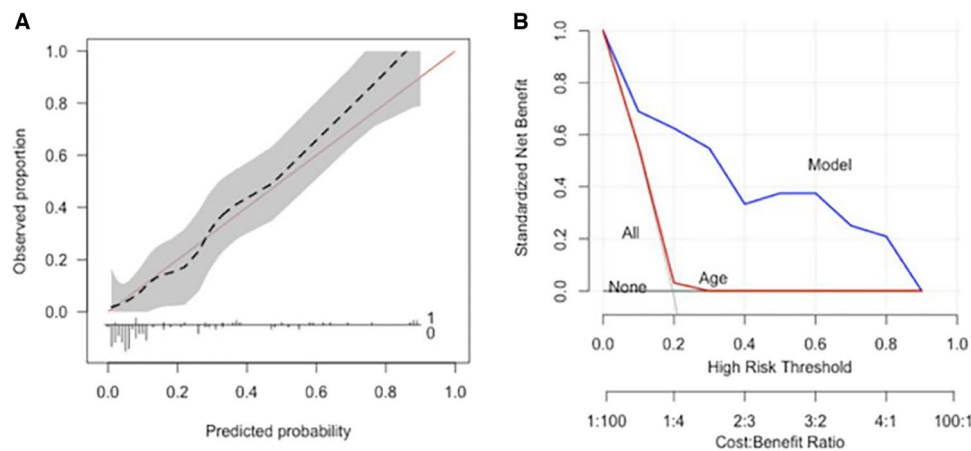
AUC area under the curve

Fig. 2 **A** Calibration plot for stochastic gradient boosting algorithm. **B** Decision curve analysis of stochastic gradient boosting algorithm. In the decision curve analysis, the net benefit of the model (blue line) relative to default strategies of changing management for all patients ("all" or for no patients ("none"). The ("all") line represents the net benefit from changing management for all patients. The line slopes down because at a threshold of zero, false positives are given no weight relative to true positives; as the threshold increases, false positives gain increased weight relative to true positives and changing management for all patients results in decreasing net benefit. The horizontal line ("none") represents the default strategy of changing management for no patients (net benefit is zero at all thresholds)

towards preoperative health optimization applications is depicted in Fig. 3.

Discussion

The main finding of the current proof-of-concept study was that the best performing machine learning algorithm conferred good predictive capability with regards to risk of experiencing a complication after primary THA at the senior author's institution. This model incorporated various modifiable and patient-specific risk factors that can be the target of optimization during the preoperative period that may decrease the risk of complications prior to undergoing surgical intervention. Furthermore, the development of a clinical decision-making tool which depicts and calculates patient-specific risk for complications can theoretically augment patient care by providing teachable and real-time data in clinic settings which may be used towards health optimization purposes. Such data may also be critical for establishing tiers for alternative payment models so as not to promote the potential for "cherry-picking" behavior and access to care problems.

There are several limitations that should be considered within the context of the current study results. Although the current machine learning algorithms had good prediction capabilities, the incidence of complications was not high enough to perform prediction of individual complication categories such as periprosthetic joint infection or myocardial infarction. The definition of complications was also broad as to capture a wide variety of potential postoperative events, which may have inflated the complication rate. However, we believe that the included complications represent primarily

major events which would be relevant to associated penalties in alternative payment models. This study is retrospective in design and therefore is subject to biases inherent in such data collection methods; however, there was a high completion rate of included data and multiple imputation methods were employed to mitigate the effect of missing data. Finally, the relatively small sample size limited the holdout (testing) dataset for assessment of model performance to only 123 patients. As such, the present study represents a proof-of-concept design and the open-access tool presented herein is merely for educational purposes until rigorous external validation is performed. It is yet to be known if the same predictive value in assessing risk will hold in other patient populations. Finally, this analysis is limited in that though we investigated the predictive importance of all variables available in this specific institutional repository, other variables important for predicting complications likely exist. Future studies are warranted that incorporate other clinically relevant variables into the current model to determine whether or not these variables confer beneficial changes to the overall performance of the model in predicting all-cause complications.

The current study demonstrated that the stochastic gradient boosting algorithm was the best performing machine learning algorithm of the five that were developed and internally validated. This particular model had an AUC of 0.88, which is considered good discriminatory capability, and demonstrated appropriate predictive probabilities relative to observed events as the model did not overfit the data (Fig. 2A). This is particularly important in reference to standard AUC values, as in real-life practice complications are better described at the patient-level as probabilities of experiencing an event as opposed to a binary all-or-nothing
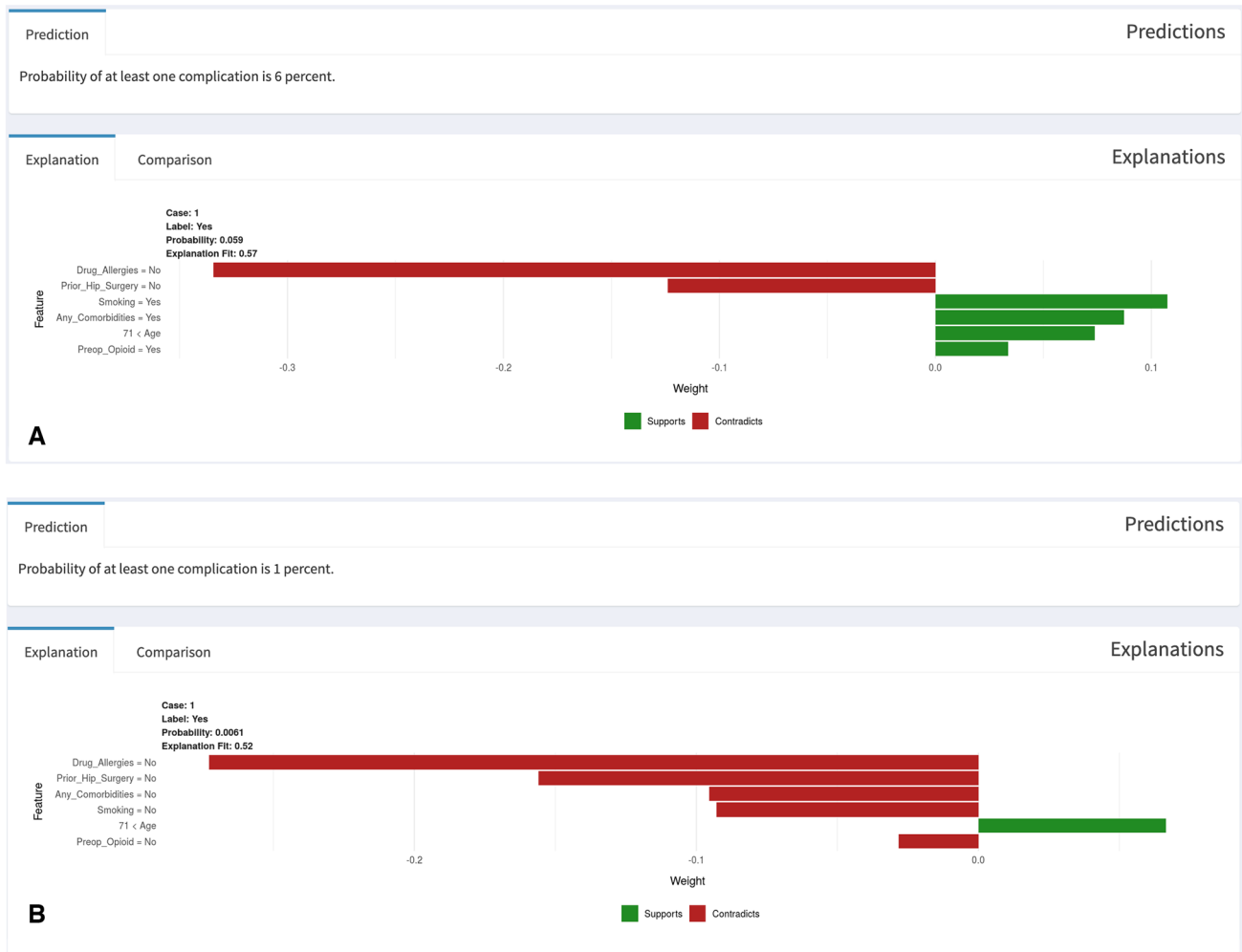
Fig. 3 Example of individual patient-level explanation for prediction of postoperative complications by stochastic gradient boosting algorithm. **A** Scenario 1 depicts a patient who is 80-years-old patient and presents to a total joint arthroplasty clinic with no prior history of hip surgery, more than one medical comorbidity, has used opioids for greater than three months to manage their current hip pain, and is a current smoker. For this particular patient, their risk of experiencing a complication after primary THA is 6.0%. **B** In scenario two, following a period of preoperative health optimization in which they quit smoking, stopped using opioid medications, and worked with a primary physician to improve their medical comorbidities, this patient reduced their risk of a postoperative complication to 1.0%. Features in red contradict (decrease the risk) of experiencing a complication, while features in green support (increase the risk) of experiencing a complication

event. Although the random forest (Supplement 1) and elastic-net penalized logistic regression (Supplement 2) also performed well, they had inferior calibration compared to the stochastic boosting gradient. Furthermore, the decision-curve analysis in the current study (Fig. 2B) demonstrated that using the stochastic gradient boosting model for risk stratification of postoperative complications was superior to considering all patients as high risk, none of the patients as high risk, and when considering age alone as a risk factor. Put simply, for patients undergoing primary THA, the stochastic gradient boosting model conferred greater utility in terms of preoperative risk stratification in comparison to alternate strategies of determining complication risk. This model provides concise assessment of complication risk in

patients undergoing primary THA based on synthesized preoperative patient data. In addition, the model was incorporated into a proof-of-concept application that is user friendly and patient-specific while requiring fewer variables than prior prediction risk models [13, 14].

The current study found that the most important patient-specific factors contributing to complications in the institutional data set under consideration were age, medication allergies, opioid use, smoking, comorbidities, and prior hip surgery. It is of note, however, that the primary outcome in this study was all-cause complications. Using this all-encompassing definition of complications was necessary as the incidence of individual complications, such as myocardial infarction, was too infrequent to develop a meaningful

model to predict each specific complication. In this context, some of the model interpretability is lost. For example, though a greater patient age may be predictive of all-cause complications, it is possible that increased age was protective of some specific complications simultaneously, such as dislocation. Few studies have also used this all-encompassing definition of complications in attempting to determine associations between preoperative variables and complications after THA. Harris et al. [36] used the American College of Surgeons-National Surgical Quality Improvement Program (ACS-NSQIP) database and least absolute shrinkage and selection operator (LASSO) methods to predict 30-day complications and mortality after total hip and knee arthroplasty procedures. The authors also found that age and various comorbidities were important contributors to experiencing complications after total knee or total hip arthroplasty. However, limitations to their study include: (1) combining total hip and knee arthroplasty patients, which are representative of potentially different patients with distinct risk profiles; (2) using LASSO methodology with less than excellent discriminatory capabilities (all models with AUCs less than 0.8, and for all complications, equal to 0.68); and (3) utilization of a national database with inherent limitations such as overrepresentation of specific populations and the limitation of 30-days complication data. Although the current study was not able to externally validate the best performing algorithm as Harris et al. did on the Veterans Affairs Surgical Quality Improvement Program (VASQIP) database, the model in the current study has the benefits of (1) using institutional data from a single-surgeon; (2) capturing complications within two-years of primary THA; and (3) having rigorously tested five independent classification-type machine learning models. Nonetheless, there remain limitations to both the current study and that performed by Harris et al. [36] which will need to be improved upon prior to creating a meaningful tool amenable to confidently predicting complications in diverse populations.

The model in the current study provides a rapid method for combining various pertinent clinical data points to accurately quantify patient risk at the individual level. Although risk stratification has been previously investigated in elective THA, research has primarily focused on global assessment of risk. The novelty of the methodology developed in the present study is the ability to efficiently determine risk at the individual patient-level and receive real-time feedback on the patient factors influencing the risk calculation. As the dataset in the current study is small and requires external validation, the presented patient scenario (Fig. 3) represents a proof-of-concept for machine learning capabilities and how machine learning can potentially impact clinical workflow and patient outcomes in the future. Though the clinical utility of this algorithm remains questionable, it is important to demonstrate how the machine learning algorithm and online application function. Future studies are warranted to externally validate the model in the current study and determine if additional variables could be of clinical utility, as well as to determine if patients undergoing total knee arthroplasty require a separate risk model. This open-source tool is for educational purposes only and should not be used in clinical settings at this time due to its generalizability being unknown outside of the authors' institution.

Conclusion

The stochastic boosting gradient algorithm demonstrated good discriminatory capacity for identifying patients at high-risk of experiencing a postoperative complication and proof-of-concept for creating office-based applications from machine learning that can perform real-time prediction. However, this clinical utility of the current algorithm is unknown and definitions of complications broad. Further investigation on larger data sets and rigorous external validation is necessary prior to the assessment of clinical utility with respect to risk-stratification of patients undergoing primary THA.

Declarations

References

1. Doran JP, Beyer AH, Bosco J, Naas PL, Parsley BS, Slover J, Zabinski SJ, Zuckerman JD, Iorio R (2016) Implementation of bundled payment initiatives for total joint arthroplasty: decreasing cost and increasing quality. Instr Course Lect 65:555–566
2. Siddiqi A, White PB, Mistry JB, Gwam CU, Nace J, Mont MA, Delanois RE (2017) Effect of bundled payments and health care reform as alternative payment models in total joint arthroplasty: a clinical review. J Arthroplasty 32(8):2590–2597
3. Kwon YM, Rossi D, MacAuliffe J, Peng Y, Arauz P (2018) Risk factors associated with early complications of revision surgery for head-neck taper corrosion in metal-on-polyethylene total hip arthroplasty. J Arthroplasty 33(10):3231–3237
4. Miettinen SS, Makinen TJ, Kostensalo I, Makela K, Huhtala H, Kettunen JS, Remes V (2016) Risk factors for intraoperative calcar fracture in cementless total hip arthroplasty. Acta Orthop 87(2):113–119

5. Triantafyllopoulos GK, Soranoglou VG, Memtsoudis SG, Sculco TP, Poultsides LA (2018) Rate and risk factors for periprosthetic joint infection among 36,494 primary total hip arthroplasties. J Arthroplasty 33(4):1166–1170

6. Shetty T, Nguyen JT, Wu A, Sasaki M, Bogner E, Burge A, Cogsil T, Kim EU, Cummings K, Su EP, Lyman S (2019) Risk factors for nerve injury after total hip arthroplasty: a case-control study. J Arthroplasty 34(1):151–156

7. Inneh IA, Lewis CG, Schutzer SF (2014) Focused risk analysis: regression model based on 5314 total hip and knee arthroplasty patients from a single institution. J Arthroplasty 29(10):2031–2035

8. Gausden EB, Parhar HS, Popper JE, Sculco PK, Rush BNM (2018) Risk factors for early dislocation following primary elective total hip arthroplasty. J Arthroplasty 33(5):1567–1571 (**e1562**)

9. Badarudeen S, Shu AC, Ong KL, Baykal D, Lau E, Malkani AL (2017) Complications after revision total hip arthroplasty in the medicare population. J Arthroplasty 32(6):1954–1958

10. Jeschke E, Citak M, Gunster C, Halder AM, Heller KD, Malzahn J, Niethard FU, Schrader P, Zacher J, Gehrke T (2018) obesity increases the risk of postoperative complications and revision rates following primary total hip arthroplasty: an analysis of 131,576 total hip arthroplasty cases. J Arthroplasty 33(7):2287–2292 (**e2281**)

11. Courtney PM, Boniello AJ, Berger RA (2017) Complications following outpatient total joint arthroplasty: an analysis of a national database. J Arthroplasty 32(5):1426–1430

12. Schilling PL, Bozic KJ (2016) Development and validation of perioperative risk-adjustment models for hip fracture repair, total hip arthroplasty, and total knee arthroplasty. J Bone Joint Surg Am 98(1):e2

13. Kunze KN, Li J, Movassaghi K, Wiggins AB, Sporer SM, Levine BR (2018) Internal validation of a predictive model for complications after total hip arthroplasty. J Arthroplasty 33(12):3759–3767

14. Oldmeadow LB, McBurney H, Robertson VJ (2003) Predicting risk of extended inpatient rehabilitation after hip or knee arthroplasty. J Arthroplasty 18(6):775–779

15. Kunze KN, Polce EM, Patel A, Courtney PM, Levine BR (2021) Validation and performance of a machine-learning derived prediction guide for total knee arthroplasty component sizing. Arch Orthop Trauma Surg 141(12):2235–2244

16. Kunze KN, Polce EM, Clapp I, Nwachukwu BU, Chahla J, Nho SJ (2021) Machine learning algorithms predict functional improvement after hip arthroscopy for femoroacetabular impingement syndrome in athletes. J Bone Joint Surg Am 103(12):1055–1062

17. Kunze KN, Karhade AV, Sadauskas AJ, Schwab JH, Levine BR (2020) Development of machine learning algorithms to predict clinically meaningful improvement for the patient-reported health state after total hip arthroplasty. J Arthroplasty 35(8):2119–2123

18. Ramkumar PN, Karnuta JM, Haeberle HS, Rodeo SA, Nwachukwu BU, Williams RJ 3rd (2021) Effect of preoperative imaging and patient factors on clinically meaningful outcomes and quality of life after osteochondral allograft transplantation: a machine learning analysis of cartilage defects of the knee. Am J Sports Med 49(8):2177–2186

19. Ramkumar PN, Karnuta JM, Haeberle HS, Owusu-Akyaw KA, Warner TS, Rodeo SA, Nwachukwu BU, Williams RJ 3rd (2021) Association between preoperative mental health and clinically meaningful outcomes after osteochondral allograft for cartilage defects of the knee: a machine learning analysis. Am J Sports Med 49(4):948–957

20. University of Wisconsin-Madison, Social Science Computing Cooperative (2013) Multiple Imputation in Stata: Deciding to Impute. Published 01/13/2013. Available at https://www.ssc.wisc.edu/sscc/pubs/stata_mi_decide.html. Accessed Mar 2021

21. Lee JH, Huber Jr, J (2011) Multiple imputation with large proportions of missing data: How much is too much? In: United Kingdom Stata Users' Group Meetings 2011. Stata Users Group, https://ideas.repec.org/p/boc/usug11/23.html. Accessed Mar 2021

22. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, Bonsel G, Badia X (2011) Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 20(10):1727–1736

23. Byrd JW (2003) Hip arthroscopy: patient assessment and indications. Instr Course Lect 52:711–719

24. van Buuren S, Groothuis-Oudshoorn K (2011) Mice: multivariate imputation by chained equations in R. J Stat Softw 45(3):1–67

25. Karhade AV, Thio Q, Ogink PT, Bono CM, Ferrone ML, Oh KS, Saylor PJ, Schoenfeld AJ, Shin JH, Harris MB, Schwab JH (2019) Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. Neurosurgery 85(4):E671–E681

26. Thio Q, Karhade AV, Ogink PT, Raskin KA, De Amorim BK, Lozano Calderon SA, Schwab JH (2018) Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma? Clin Orthop Relat Res 476(10):2040–2048

27. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21(1):128–138

28. Steyerberg EW, Vergouwe Y (2014) Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 35(29):1925–1931

29. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 26(6):565–574

30. Vickers AJ, Cronin AM, Elkin EB, Gonen M (2008) Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak 8:53

31. Brier GW, Allen RA (1951) Verification of weather forecasts. In: Malone TF (ed) Compendium of meterology. American Meteorological Society, Boston, pp 841–848

32. Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd SIGKDD international conference on knowledge discovery and data mining. pp 1135–1144. https://arxiv.org/abs/1602.04938. Accessed March 2021

33. Ribeiro MT, Singh S, Guestrin C (2019) Model-agnostic interpretability of machine learning. In: Cornell University. https://arxiv.org/abs/1606.05386

34. Collins GS, Reitsma JB, Altman DG, Moons KG (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Surg 102(3):148–158

35. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M (2016) Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 18(12):e323

36. Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori NJ (2019) 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 Relat Res 477(2):452–460