



# Predicting extended hospital stay following revision total hip arthroplasty: a machine learning model analysis based on the ACS-NSQIP database

Tony Lin-Wei Chen<sup>1,2</sup> · MohammadAmin RezazadehSaatlou<sup>1</sup> · Anirudh Buddhiraju<sup>1</sup> · Henry Hojoon Seo<sup>1</sup> · Michelle Riyo Shimizu<sup>1</sup> · Young-Min Kwon<sup>1</sup> 

Received: 5 April 2024 / Accepted: 3 September 2024

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

## Abstract

**Introduction** Prolonged length of stay (LOS) following revision total hip arthroplasty (THA) can lead to increased health-care costs, higher rates of readmission, and lower patient satisfaction. In this study, we investigated the predictive power of machine learning (ML) models for prolonged LOS after revision THA using patient data from a national-scale patient repository.

**Materials and methods** We identified 11,737 revision THA cases from the American College of Surgeons National Surgical Quality Improvement Program database from 2013 to 2020. Prolonged LOS was defined as exceeding the 75th value of all LOSs in the study cohort. We developed four ML models: artificial neural network (ANN), random forest, histogram-based gradient boosting, and k-nearest neighbor, to predict prolonged LOS after revision THA. Each model's performance was assessed during training and testing sessions in terms of discrimination, calibration, and clinical utility.

**Results** The ANN model was the most accurate with an AUC of 0.82, calibration slope of 0.90, calibration intercept of 0.02, and Brier score of 0.140 during testing, indicating the model's competency in distinguishing patients subject to prolonged LOS with minimal prediction error. All models showed clinical utility by producing net benefits in the decision curve analyses. The most significant predictors of prolonged LOS were preoperative blood tests (hematocrit, platelet count, and leukocyte count), preoperative transfusion, operation time, indications for revision THA (infection), and age.

**Conclusions** Our study demonstrated that the ML model accurately predicted prolonged LOS after revision THA. The results highlighted the importance of the indications for revision surgery in determining the risk of prolonged LOS. With the model's aid, clinicians can stratify individual patients based on key factors, improve care coordination and discharge planning for those at risk of prolonged LOS, and increase cost efficiency.

**Keywords** Revision total hip arthroplasty · Machine learning · Artificial intelligence · Hospital stay · Risk factors · Clinical decision support

## Introduction

Total hip arthroplasty (THA) is one of the most common orthopedic procedures performed in the United States, with approximately 114,000 THA procedures performed each year [1]. Despite the high success rate of THA, approximately 4.3% of patients require revision THA within 10 years [2]. Revision THA is associated with increased surgical complexity, longer operating time, and a higher risk of complications compared to primary THA [3]. The cost of revision THA is also markedly higher, with the average cost per episode of care reported to be \$87,000 in 2014 [2].

✉ Young-Min Kwon  
ymkwon@mgh.harvard.edu

<sup>1</sup> Bioengineering Laboratory, Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street, Boston, MA 02114, USA

<sup>2</sup> Department of Biomedical Engineering, The Hong Kong Polytechnic University, Yuk Choi Rd 11, 999077 Hong Kong SAR, China

The total length of stay (LOS) after revision THA is an important factor that influences both patient outcomes and hospital expenditure [4–6]. Previous studies have predicted an extra day in the hospital to increase the cost burden by \$2,000–\$3,000 [7]. The ability to predict prolonged LOS in individual patients can encourage proactive measures and allocate resources for those patients, thereby improving treatment efficiency and reducing care costs [8].

Traditional statistical models have been historically used to predict prolonged LOS following total joint arthroplasty [9–12]. Yet, the model performances were inherently limited by the model linearity and simplification of the variable interrelationship. Recently introduced machine learning (ML) models have outperformed such statistical approaches in terms of accuracy and predictive performance in a variety of contexts [13–16]. As a result, ML models are increasingly being used to predict outcomes in a clinical setting [17]. Previous studies have reported excellent ML model performance in identifying patients at high risk of prolonged LOS after primary total joint arthroplasty [18–21]. With rising interest in ML models, there has been an increased call for action to broaden the applicability of the model by incorporating multi-center patient data as a way to establish the model's generalizability while maintaining a high level of accuracy [22, 23].

Large national datasets have been recommended due to their accessibility and availability of a variety of clinical and demographic data on surgical cases across the United States. The American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) is one such database that aggregates patient data from multiple sites in the United States [15]. This study aimed to develop ML models using the ACS-NSQIP dataset and evaluate their performance across three domains: discrimination, calibration, and clinical utility for the prediction of prolonged LOS after revision THA.

## Materials and methods

### Patient cohort, variables, and study outcomes

The ACS-NSQIP databases from 2013 to 2022 were reviewed to acquire the data of patients who underwent revision THA. The CPT codes 27,134, 27,137, and 27,138 were used to identify our research population. The exclusion criteria were age under 18 or above 100, body mass index (BMI) higher than 100, emergency surgery, bilateral arthroplasty, and incomplete/unclear hospital records. Data were also excluded if a negative value was recorded in any preoperative blood test. For example, a white blood cell count of -99 thousand/mm<sup>3</sup> was considered a faulty entry

as the value was physiologically impossible. This study was reviewed and approved by the institutional internal review board. The procedure for ML modeling in this study was reported following an established publication guideline [24].

The prediction target of the ML models was prolonged LOS following revision THA. In line with prior research [10, 25], we used the 75th percentile of all LOSs (3 days) as the cut-off value to divide the study cohort into two classes: normal LOS and prolonged LOS. Inputs to the ML models included sociodemographic variables (age, sex, race, ethnicity, BMI, and smoking history), comorbidities (dyspnea, diabetes, hypertension, bleeding disorders, congestive heart failure, etc.), and perioperative variables (American Society of Anesthesiologists (ASA) score, blood test results, total operation time, blood loss, transfusion, anesthesia type, etc.).

### Model development and performance analysis

The study cohort was split into training and testing datasets at a ratio of 8:2 using the stratified slip technique [26]. Continuous variables were standardized and no imputation was performed for unknown feature values. ML models included in the study were: artificial neural network (ANN), random forest (RF), histogram-based gradient boosting (HGB), and k-nearest neighbors (KNN). These models were selected based on their previous performance on similar classification tasks [18, 26]. We applied recursive feature elimination using a rudimentary RF model (constructed by passing default values of the hyperparameters) to streamline the feature list while maintaining the model's discrimination capacity, which was indicated by the area under the receiver operating curve (AUC). All patient variables were first fit to the model and ranked based on their contribution to the prediction, afterward the least important variable was removed. The procedure was repeated until the last variable remained. The model's performance in each iteration was recorded to determine the optimal number of patient variables. We found that the model's AUC value was steady at the beginning of elimination and gradually dropped after the number of remaining features was lower than 30. Therefore, the top 30 important patient variables were reserved for the subsequent model development. Important hyperparameters of each ML model were tuned using a coarse-grained grid-search method. In brief, the value of each hyperparameter was allowed to vary within a predefined range based on our previous studies on similar topics [23, 27]. A "grid" was comprised of dots that represented all possible combinations of the hyperparameter set. A "coarse-grained" method tested a subset of the dots and identified the combination that produced the best prediction accuracy of the model.

The hyperparameters and their corresponding ranges for each ML model are as follows: ANN: learning rate: 0.0001, 0.001 ... 0.1, the size of the hidden layer: 3 to 5, the number of neurons: 50 to 100 for each layer, and maximum epochs: 30; RF: the number of trees: 5 to 100 at the interval of 5, and minimal sample leaf: 2 to 50 at the interval of 2; HGB: learning rate: 0.0001, 0.001 ... 0.1, the maximum number of interaction: 60 to 160 at the interval of 20, and leaves: 21 to 46 at the interval of 5; KNN: the number of neighbors: 50 to 450 at the interval of 50, and distance metrics: weights: uniform or distance, and  $p$ : 1 or 2. Fivefold cross-validation with five repetitions was applied during model development. The training dataset was divided into five subsets, with each subset serving as a validation set once while the remaining four were used for training. This process was repeated five times with different data splits. The repetition of cross-validation was implemented following the final feature selection to mitigate the risk of overfitting and reduce variances in the models' performance metrics. After the training was completed, the models were applied to the testing dataset. The average computing time to develop the ML models ranged from 72 to 434 s on a computer running Microsoft Windows 10 Pro (Microsoft Corp., Redmond, Washington, USA), equipped with an Intel i7-13700 F CPU (Intel Corp., Santa Clara, California, USA), an NVIDIA GeForce RTX 3060 GPU (NVIDIA Corp., Santa Clara, California, USA), and 32 GB RAM.

The model's performance was assessed using several metrics. The first metric was AUC which determines the model's discrimination. An AUC value greater than 0.80 indicates that a model has excellent discrimination [22, 23]. The second metric used was calibration plots, which graphically represent the agreement between the actual outcomes and the model-predicted probability. A well-calibrated model has a slope of 1 and an intercept of 0. The third metric used was the Brier score, which measures the mean squared difference between the predicted probabilities and the actual outcomes of an event. A Brier score approximating 0 indicates that a model has few prediction errors [28]. Lastly, the decision curve analysis was used to evaluate the benefit of using the model compared to treating all or none of the patients across a range of probabilities [29]. The model's interpretability was explained globally and locally. The plot of feature importance identified the patient factors with the greatest influence on the model prediction, while a local explanation was provided for a representative patient to demonstrate the weight of each variable on the final prediction of the machine learning model. Codes for ML modeling and computing performance metrics are accessible at <https://github.com/tlwchen/ML-models-for-event-prediction>.

We anticipated that there may be differences in sex ratio between the normal LOS and prolonged LOS groups. As

several measures of the patient characteristics, such as the hematocrit level, can vary in females compared to males, skewness in data distribution might bias the model performance. We therefore stratified the study cohort by sex and performed secondary modeling for each subgroup using the ML model that demonstrated the best predictive metrics for prolonged LOS. The model performance was then compared between the subgroups to ascertain the conjecture of sex-specific influences on prediction accuracy.

## Data analysis

Baseline patient characteristics between the normal LOS and prolonged LOS groups were compared. Continuous variables were analyzed using either the independent student T-test or the Mann-Whitney U-test, contingent on whether the assumptions of parametric tests were violated. The Chi-square test was utilized to examine nominal variables. Cohen's  $d$ , rank-biserial correlation coefficient, and Cramér's  $V$  were calculated to indicate effect sizes for corresponding statistical models used in primary examinations. Effect sizes were interpreted using the Cohen convention of negligible ( $<0.20$ ), small ( $0.20$ – $0.49$ ), medium ( $0.50$ – $0.79$ ), and large ( $>0.80$ ) values [30]. Statistical analyses were carried out utilizing Anaconda (version 2.5.4, Anaconda Inc., Austin, TX, USA), Python (version 3.11.4, Python Software Foundation, Wilmington, DE, USA), and SPSS (version 18.0, IBM Corp., Armonk, NY, USA).  $P < 0.05$  was considered for the level of statistical significance.

## Results

### Patient characteristics

A total of 11,749 patients were included in the analysis, of which 26.8% had extended LOS ( $N=3153$ ). The percentage of male patients was slightly higher in the normal LOS group than in the prolonged LOS group (45.53% vs. 44.21%,  $p < 0.001$ ). Statistics showed that patients with prolonged LOS were older (68.55 years vs. 65.92 years,  $p < 0.001$ ), had a higher ASA score (ASA level 3 or above: 78.64% vs. 57.94%,  $p < 0.001$ ) and comorbidity rates (hypertension, COPD, diabetes, etc.  $p < 0.005$ ) compared to those with normal LOS (Table 1). A greater percentage of patients in the prolonged LOS group were smokers (17.69% vs. 13.91%,  $p < 0.001$ ) and ethnic minorities (12.98% vs. 11.35%,  $p = 0.07$ ). Patients from the prolonged LOS group also presented suboptimal blood test results (higher leukocyte counts and reduced hematocrit,  $p < 0.001$ ), had longer total operation time (174.62 min vs. 136.47 min,  $p < 0.001$ ),

**Table 1** Baseline characteristics of patients undergoing revision total hip arthroplasty in the study cohort

Feature	Extended LOS (N: 3153)	Normal LOS (N: 8596)	P-value	Effect size
<b>Sociodemographic</b>				
Age (years)	68.55 ± 12.53 (68.11—68.99)	65.92 ± 11.53 (66.16—65.67)	<0.001	0.10
BMI (Kg/m <sup>2</sup> )	29.94 ± 7.54 (29.68—30.21)	30.00 ± 6.61 (29.87—30.15)	<0.001	0.02
Sex (male, %)	44.21 (1394)	45.53 (3914)	0.200	0.01
Race (white, %)	88.65 (2795)	87.02 (7480)	0.070	0.03
Hispanic (%)	2.94 (93)	2.67 (230)	0.440	0.04
ASA score			<0.001	0.10
	Level 1 (%)	1.30 (112)		
	Level 2 (%)	40.75 (3503)		
	Level 3 (%)	55.14 (4740)		
	Level 4 (%)	2.80 (241)		
Indication for revision surgery			<0.001	0.16
	Loosening/wear (%)	82.89 (7125)		
	Infection (%)	9.67 (831)		
	Instability (%)	0.50 (43)		
	Stiffness (%)	6.95 (597)		
<b>Laboratory tests</b>				
Leukocyte count (thousands/mm <sup>3</sup> )	7.11 ± 2.47 (7.06—7.16)	7.90 ± 3.07 (7.80—8.01)	<0.001	0.12
Hematocrit (%)	35.40 ± 5.67 (35.20—35.60)	39.85 ± 4.84 (39.75—39.95)	<0.001	0.35
Platelet count (thousands/mm <sup>3</sup> )	257.75 ± 102.87 (254.16—261.34)	252.91 ± 78.83 (251.25—254.58)	<0.001	0.01
<b>Comorbidities</b>				
Smoking (%)	17.69 (558)	13.91 (1196)	<0.001	<0.01
Dyspnea (%)			<0.001	0.02
	At Rest (%)	0.29 (25)		
	Moderate (%)	4.43 (381)		
	None (%)	95.28 (8190)		
Severe COPD (%)	8.91 (281)	5.17 (444)	<0.001	0.02
Ascites (%)	0.22 (7)	0.03 (3)	0.005	0.02
Congestive heart failure (%)	2.15 (68)	0.52 (45)	<0.001	0.03
Hypertension (%)	66.25 (2089)	58.84 (5058)	<0.001	0.01
Dialysis (%)	1.77 (56)	0.46 (40)	<0.001	0.02
Disseminated cancer (%)	1.55 (49)	0.41 (35)	<0.001	0.04
Wound infection (%)	6.85 (216)	11.40 (980)	0.002	0.07
Preoperative steroid Use (%)	7.86 (248)	6.22 (535)	0.002	0.03
Diabetes	18.07 (570)	13.18 (1133)	<0.001	0.02
Preoperative transfusion (%)	4.19 (132)	0.32 (28)	<0.001	0.06
<b>Intra-operation parameters</b>				
Operation time (min)	174.62 ± 84.48 (171.67—177.57)	136.47 ± 63.79 (135.13—137.82)	<0.001	0.22
Wound classification			<0.001	0.10
	Clean (%)	92.25 (7930)		
	Clean/contaminated (%)	2.07 (178)		
	Contaminated (%)	1.52 (131)		
	Dirty/infected (%)	4.15 (357)		
Anaesthesia			<0.001	0.09
	General (%)	69.45 (5970)		
	Spinal (%)	16.97 (1459)		
	MAC/IV sedation (%)	11.47 (986)		
	Regional (%)	1.58 (136)		
	Others (%)	0.52 (45)		

**Table 1** (continued)

Feature	Extended LOS (N: 3153)	Normal LOS (N: 8596)	P-value	Effect size
<b>Hospitalization</b>				
Total length of stay (days)	8.29 ± 5.05 (8.11–8.47)	2.33 ± 1.02 (2.31–2.35)	< 0.001	

Continuous variables are presented as mean ± standard deviation (95% CI). Categorical variables were presented as percentage occupied (raw number). BMI: body mass index; ASA: American Society of Anesthesiologists; COPD: chronic obstructive pulmonary disease; MAC/IV: monitored anesthesia care/intravenous

**Table 2** Discrimination and calibration performance of the machine learning models for predicting LOS during training

Metric	ANN	RF	HGB	KNN
AUC	0.84 (0.82–0.86)	0.82 (0.80–0.83)	0.82 (0.81–0.82)	0.82 (0.81–0.85)
Calibration (slope)	0.96 (0.90–1.03)	1.03 (0.98–1.06)	0.91 (0.86–1.02)	1.36 (1.22–1.57)
Calibration (intercept)	0.02 (0.01–0.04)	-0.01 (-0.03–0.01)	0.03 (-0.02–0.05)	0.01 (-0.01–0.05)
Brier score	0.136 (0.123–0.143)	0.142 (0.139–0.147)	0.143 (0.137–0.147)	0.153 (0.147–0.158)

ANN: artificial neural network; RF: random forest; HGB: histogram-based gradient boosting; KNN: k-nearest neighbor; AUC: area under the receiver operating characteristic curve. An AUC value higher than 0.8 indicates an excellent discrimination capacity of a machine learning model. A calibration slope and calibration intercept equal to 1 and 0 respectively indicate an ideally calibrated model, by which the predicted values completely match the actual outcomes. A Brier score approximating 0 indicates that a model has few prediction errors

**Table 3** Discrimination and calibration performance of the machine learning models for predicting LOS during validation

Metric	ANN	RF	HGB	KNN	ANN (secondary modeling)	
					For females	For males
AUC	0.82	0.83	0.83	0.83	0.80	0.84
Calibration (slope)	0.90	1.03	0.89	1.32	0.88	0.87
Calibration (intercept)	0.02	-0.01	0.03	0.02	0.05	0.06
Brier score	0.140	0.140	0.139	0.151	0.150	0.133

ANN: artificial neural network; RF: random forest; HGB: histogram-based gradient boosting; KNN: k-nearest neighbor; AUC: area under the receiver operating characteristic curve. The second to fifth columns display the performance metrics of the included machine learning models during the testing session. The sixth and seventh columns show ANN's performance during secondary modeling for the female and male subgroups respectively. An AUC value higher than 0.8 indicates an excellent discrimination capacity of a machine learning model. A calibration slope and calibration intercept equal to 1 and 0 respectively indicate an ideally calibrated model, by which the predicted values completely match the actual outcomes. A Brier score approximating 0 indicates that a model has few prediction errors

and were more likely to receive transfusion before surgery (Table 1). A larger fraction of the revision surgeries were caused by infection (33.11% vs. 9.31%,  $p < 0.001$ ) in the prolonged LOS group. Despite the statistical significance, the effect sizes of the differences between the two patient groups were generally small across the patient variables (Table 1).

### Assessment of model performance

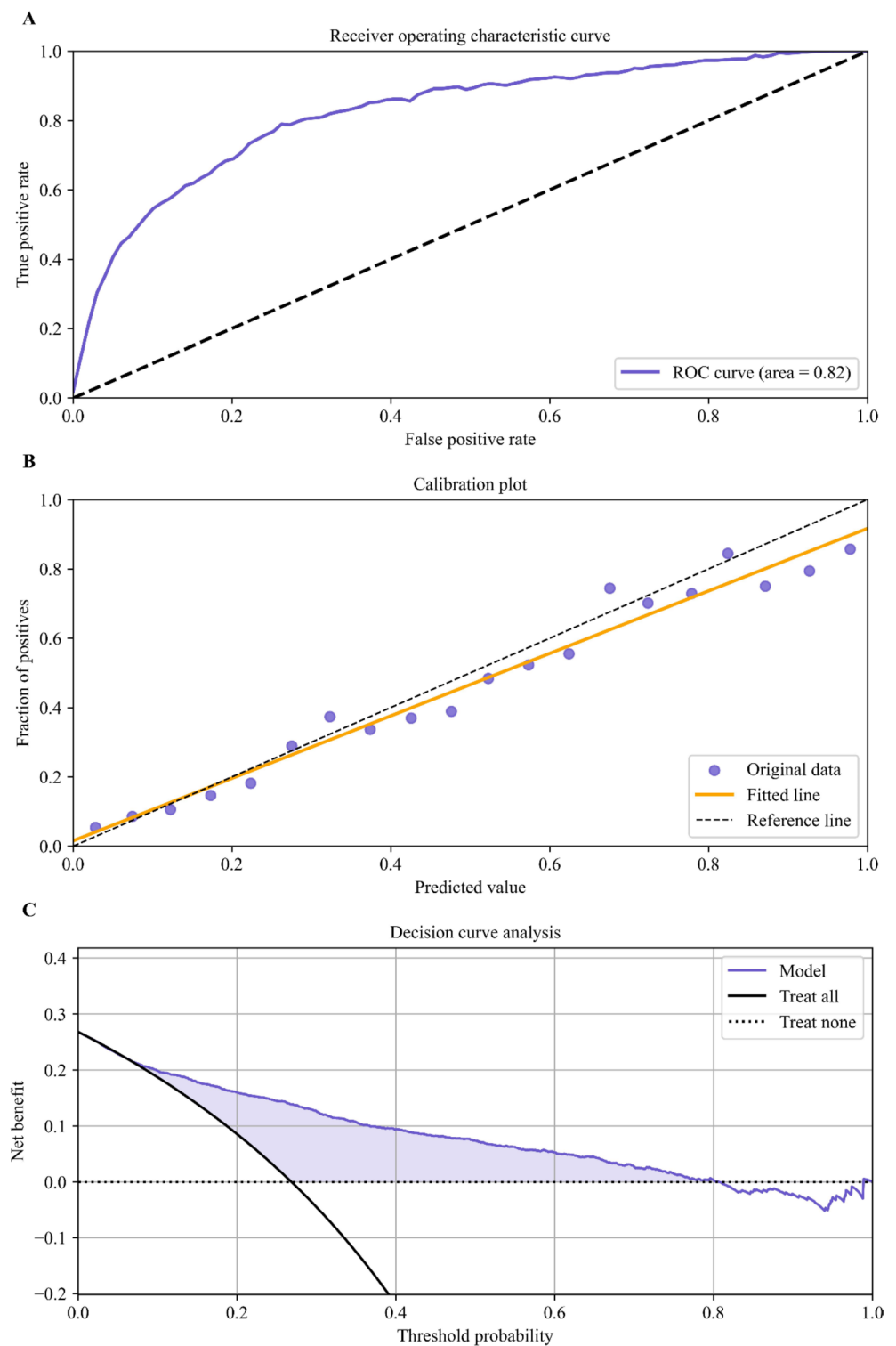
All models showed excellent discrimination and calibration performance in the training session. The five-fold cross-validation for all models reported an AUC of 0.83 to 0.88, a calibration slope of 0.84 to 1.32, a calibration intercept of -0.08 to 0.03, and a Brier score of 0.087 to 0.132 (Table 2). A similar level of performance was retained across the models with the testing dataset (Table 3). ANN delivered the

best results in predicting prolonged LOS after revision THA (AUC: 0.82, calibration slope: 0.90, calibration intercept: 0.02, Brier score: 0.140, Fig. 1A–B). The decision curve analysis demonstrated that ANN produced higher net benefits than the default strategies that assumed all or no observations were positive (Fig. 1C). ANN was therefore selected for secondary modeling for each sex subgroup. Following a similar modeling procedure, the results showed comparable performance metrics of ANN between sexes (Table 3).

### Model interpretation

The plot of feature importance revealed that the patient factors that most contribute to prolonged LOS after revision THA were preoperative blood tests (hematocrit  $< 36.81\%$ , platelet count  $> 253.54$  thousand/mm<sup>3</sup>, and white blood cell count  $> 7.51$  thousand/mm<sup>3</sup>), preoperative transfusion,

**Fig. 1** Plots of performance metrics for artificial neural network. **(A)** the receiver operating characteristics curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) of the model prediction at different classification thresholds. With a lower threshold, more items will be classified as positive, which increases both TPR and FPR. The area under the curve (AUC) is an aggregate measure of the model performance across all possible classification thresholds. A good model that is able to effectively distinguish between positive cases and negative cases will generate high TPR/FPR ratios at any threshold and therefore produce a high AUC value. **(B)** the calibration plot shows the agreement between the model predictions and observations in different percentiles of the predicted values. A complete match between predictions and observations will generate a diagonal line (slope: 1, intercept: 0). **(C)** the decision curve analysis compares the net benefit (a trade-off between the benefits and harms of a particular decision) of using a predictive model to those of two baseline strategies: “treating all” and “treating none” at different probability thresholds. A model showing higher net benefit than the baseline strategies possesses clinical utility



operation time ( $> 135.26$  min), indications for revision THA (infection), and age ( $> 76$  years) (Fig. 2). A local explanation of an individual who stayed 2 days following revision THA featured a male aged 65 years with an ASA level of 3 and normal laboratory test results. The patient required transfusion before surgery and underwent revision THA for 145 min. ANN predicted that his probability of experiencing prolonged LOS was 38.33%.

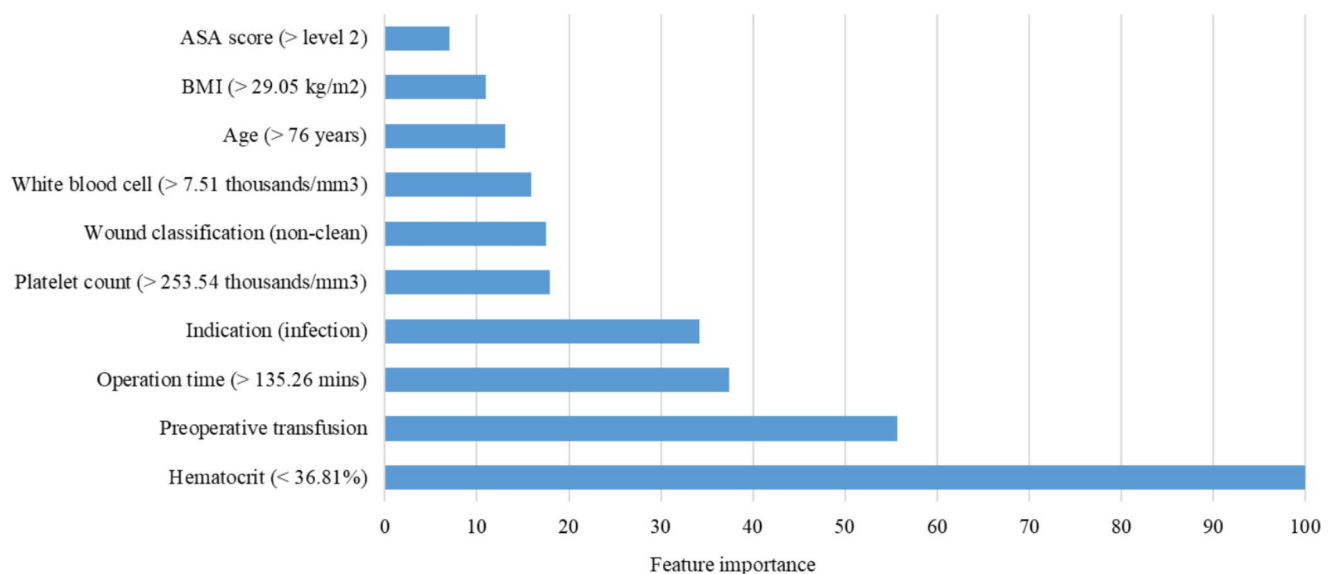
## Discussion

The study developed four ML models to predict prolonged LOS using a comprehensive national dataset containing over 11,749 revision THA patients. Our findings indicated that all models had great prediction performance during training and testing sessions. ANN provided the most accurate predictions of patients subject to lengthened hospital stays, as reflected by its high level of AUC, calibration parameters, and Brier score. ANN also showed clinical utility by producing net benefits against varying probability thresholds in the decision curve analysis. Important predictors of prolonged LOS, as indicated by ANN, were preoperative laboratory results, preoperative transfusion, operation time, indications for the revision surgery, and age.

The application of ML models to predict complications after revision arthroplasty is a relatively recent field of research [31]. Our findings were similar to those of a previous study [32] based on a single institution database, which reported an AUC of 0.84–0.87, calibration slope of 0.85–1.12, and calibration intercept of 0.14–0.21 for ML models that predicted LOS after revision total knee arthroplasty. These results of previous studies and the current study

indicated that the ML model was able to retain a high-level performance across different arthroplasty types as it consistently required several key variables to predict LOS, such as age, BMI, operation time, ASA score, and comorbidities [33]. In a retrospective study including 1,278 patients undergoing primary THA, Farley et al. [34] found that increasing age, high BMI, and comorbidities contributed to increased LOS. Roger et al. [35] identified older age, high ASA score, comorbidities, and long operation time as the risk factors of prolonged LOS following total joint arthroplasties. In addition to these previously established determinants, our models also highlighted the role of laboratory tests (total leukocyte count, hematocrit, and platelet count) in making accurate predictions, which was consistent with a previous report on using ML models to identify patients predisposed to prolonged LOS after primary THA [26]. Despite the correlation between sex and the level of these blood biomarkers, the result of secondary modeling in our study did not support the influence of sex-specific data skewness on the ML model's performance. The reliance of the model prediction on an isolated factor appeared to be insensitive to sex types. Another explanation may be that the difference in sex ratio between the two LOS groups was small, therefore limited class imbalance effects were introduced during model development to bias the model decision.

Our study also found indication for revision TKA to be an important determinant of LOS after surgery, with infection being a significant contributor to prolonged LOS. This finding is in concordance with reports by Klemm et al. [32]. Periprosthetic infection is one of the most commonly seen indications for revision TKA and oftentimes results in patient dissatisfaction and poor surgical outcomes [36]. Infection is also associated with increased complexity of



**Fig. 2** Global feature importance plot of the machine learning models for predicting prolonged length of stay after revision total hip arthroplasty

the revision surgery and a higher risk of postoperative complications [37–39], which entails extended monitoring and care support during recovery [36]. This finding substantiated the clinical utility of the ML models in decision curve analyses as it provided useful information for pre-operation patient counseling based on the type of indications for revision THA. It is worth noting that the small effect sizes of between-group differences at baseline should not be used equivalently to interpret the clinical significance of the model's prediction. The risk of prolonged LOS on the individual level is not solely reliant on the value of isolated patient features. Our results underscore the ML model's strength in detecting the hidden pattern across various data domains and deriving predictions from the collective effects of multiple pertinent factors. This advantage persists when comparing ML models to conventional logistic regression analysis. ML models excel at capturing complex interactions among variables without assuming linearity between the predictors and outcomes, which is the premise for regression analyses but usually does not hold true in high-dimension data. Although logistic regression offers better interpretability due to its simplified model structure, it is less likely to outperform ML models in predicting prolonged LOS given the large scale and complexity of the dataset in this study.

As indicated by the ML models in our study, the list of important predictors of prolonged LOS included a combination of both modifiable and unmodifiable patient factors. The components of the laboratory tests are modifiable factors that had a major contribution to the model prediction. Various clinical management and supplement strategies are available to optimize the number of white blood cell counts and hematocrit levels before surgery, thereby mitigating the risk of prolonged LOS [40]. For instance, increased leukocyte count can be addressed by identifying the underlying infection and treatments through target antibiotic therapies [41]. Preoperative screening for infections allows timely intervention, which is crucial in preventing postoperative complications [42]. Chronic conditions such as diabetes are also potential causes of abnormal white blood cell counts [43]. Hyperglycemia can impair leukocyte function, increasing the susceptibility to infections and prolonging hospital stays. Effective glycemic control through adjustments in diabetic medication and lifestyle modifications, including dietary changes, can improve immune function and the white blood cell level. Unmodifiable predictors, on the other hand, underlay the applicability of the ML models by consolidating the accuracy and reliability in identifying individuals at risk of extended hospitalization. Incorporating the models into clinical routine has the potential to allow clinicians to better stratify patients based on their risk profiles. This stratification can facilitate preoperative counseling regarding expectations of potential hospital arrangements

and more importantly, inform discharge planning, ensuring that appropriate resources and support are available for patients at risk to reduce costs associated with delayed transitions and patient dissatisfaction. The positive net benefits produced by ML models in decision curve analysis also support their clinical utility in helping with decision-making.

There are several limitations to consider when interpreting the outcomes of this study. First, the study was based on retrospective data from the ACS-NSQIP database. The accuracy of the model prediction might be affected by selection bias and data misrepresentation during manual information entry. The study cohort only included patients who had revision THA between 2013 and 2020, the findings may not be applicable in a different timeline. Second, we used a dichotomization method to categorize the outcome of the hospital stay. This method facilitates discrete class labeling in ML modeling but has the limitation of reducing the data granularity and discarding the nuanced information of variability in an LOS spectrum. The cut-off method of using the 75th percentile value may not generalize to other clinical settings or patient populations due to possibly different data distribution. This threshold is empirical and does not necessarily reflect the latest clinically meaningful distinctions. Finally, the ML models were created using a limited number of patient characteristics. Other patient factors, such as the surgical approach and ambulation protocol, have been previously reported to influence LOS but were not included in our study as they were not recorded in the ACS-NSQIP database. The actual benefits of the ML models in predicting patient outcomes warrant further investigation.

In conclusion, this study utilized a national-scale patient cohort to develop ML models that accurately predicted prolonged LOS following revision THA. ANN yielded the best performance in outcome discrimination and calibration during both training and testing sessions. All models showed great clinical utility in the decision curve analyses. Important predictors of LOS included preoperative laboratory tests, preoperative transfusion, operation time, indications for the revision surgery, and age. The integration of ML models into clinical workflows may assist in optimizing patient-specific care coordination, discharge planning, and cost containment after revision THA surgery.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00402-024-05542-9>.

## Declarations

**Completing interest** The authors have no relevant financial or non-financial interests to disclose.



## References

- American Joint Replacement Registry (AJRR) (2022) 2022 Annual Report. Rosemont, IL: American Academy of Orthopaedic Surgeons (AAOS)
- Deere K, Whitehouse MR, Kunutsor SK, Sayers A, Mason J, Blom AW (2022) How long do revised and multiply revised hip replacements last? A retrospective observational study of the National Joint Registry. *Lancet Rheumatol* 4:e468–e479. [https://doi.org/10.1016/S2665-9913\(22\)00097-2](https://doi.org/10.1016/S2665-9913(22)00097-2)
- Schwartz BE, Piponov HI, Helder CW, Mayers WF, Gonzalez MH (2016) Revision total hip arthroplasty in the United States: national trends and in-hospital outcomes. *Int Orthop* 40:1793–1802. <https://doi.org/10.1007/s00264-016-3121-7>
- Benito J, Stafford J, Judd H, Ng M, Corces A, Roche MW (2022) Length of Stay increases 90-day Readmission Rates in patients undergoing primary total joint arthroplasty. *JAAOS Glob Res Rev* 6. <https://doi.org/10.5435/JAAOSGlobal-D-21-00271>
- Burn E, Edwards CJ, Murray DW, Silman A, Cooper C, Arden NK et al (2018) Trends and determinants of length of stay and hospital reimbursement following knee and hip replacement: evidence from linked primary care and NHS hospital records from 1997 to 2014. *BMJ Open* 8:e019146. <https://doi.org/10.1136/bmjopen-2017-019146>
- Culler SD, Jevsevar DS, Shea KG, Wright KK, Simon AW (2015) The Incremental Hospital cost and length-of-stay Associated with treating adverse events among Medicare beneficiaries undergoing TKA. *J Arthroplasty* 30:19–25. <https://doi.org/10.1016/j.arth.2014.08.023>
- Schwartz AJ, Clarke HD, Sassoon A, Neville MR, Etzioni DA (2020) The clinical and financial consequences of the centers for Medicare and Medicaid Services' two-midnight rule in total joint arthroplasty. *J Arthroplasty* 35:1–6e1. <https://doi.org/10.1016/j.arth.2019.08.048>
- Zhong H, Poeran J, Gu A, Wilson LA, Gonzalez Della Valle A, Memtsoudis SG et al (2021) Machine learning approaches in predicting ambulatory same day discharge patients after total hip arthroplasty. *Reg Anesth Pain Med* 46:779–783. <https://doi.org/10.1136/rapm-2021-102715>
- Ding Z, Xu B, Liang Z, Wang H, Luo Z, Zhou Z (2020) Limited Influence of Comorbidities on length of stay after total hip arthroplasty: experience of enhanced recovery after surgery. *Orthop Surg* 12:153–161. <https://doi.org/10.1111/os.12600>
- Inneh IA, Iorio R, Slover JD, Bosco JA (2015) Role of Sociodemographic, co-morbid and Intraoperative Factors in length of Stay following primary total hip arthroplasty. *J Arthroplasty* 30:2092–2097. <https://doi.org/10.1016/j.arth.2015.06.054>
- Papalia R, Zampogna B, Torre G, Papalia GF, Vorini F, Bravi M et al (2021) Preoperative and Perioperative Predictors of Length of Hospital Stay after primary total hip arthroplasty—our experience on 743 cases. *J Clin Med* 10:5053. <https://doi.org/10.3390/jcm10215053>
- Rudasill SE, Dattilo JR, Liu J, Nelson CL, Kamath AF (2018) Do illness rating systems predict discharge location, length of stay, and cost after total hip arthroplasty? *Arthroplasty Today* 4:210–215. <https://doi.org/10.1016/j.artd.2018.01.004>
- Aram P, Trela-Larsen L, Sayers A, Hills AF, Blom AW, McCloskey EV et al (2018) Estimating an Individual's probability of revision surgery after knee replacement: a comparison of modeling approaches using a National Data Set. *Am J Epidemiol* 187:2252–2262. <https://doi.org/10.1093/aje/kwy121>
- Karnuta JM, Luu BC, Haeberle HS, Saluan PM, Frangiamore SJ, Stearns KL et al (2020) Machine Learning Outperforms Regression Analysis To Predict Next-Season Major League Baseball Player Injuries: Epidemiology and Validation of 13,982 player-years from performance and Injury Profile trends, 2000–2017. *Orthop J Sports Med* 8:232596712096304. <https://doi.org/10.1177/2325967120963046>
- Liu Y, Ko CY, Hall BL, Cohen ME ACS NSQIP Risk Calculator Accuracy using a machine learning Algorithm compared to regression. *J Am Coll Surg* 2023;Publish Ahead of Print. <https://doi.org/10.1097/XCS.0000000000000556>
- Shah AA, Devana SK, Lee C, Bugarin A, Lord EL, Shamie AN et al (2021) Prediction of major complications and readmission after lumbar spinal Fusion: a machine learning–Driven Approach. *World Neurosurg* 152:e227–e234. <https://doi.org/10.1016/j.wneu.2021.05.080>
- Lopez CD, Gazgalis A, Boddapati V, Shah RP, Cooper HJ, Geller JA (2021) Artificial Learning and Machine Learning Decision Guidance Applications in total hip and knee arthroplasty: a systematic review. *Arthroplasty Today* 11:103–112. <https://doi.org/10.1016/j.artd.2021.07.012>
- Abbas A, Mosseri J, Lex JR, Toor J, Ravi B, Khalil EB et al (2022) Machine learning using preoperative patient factors can predict duration of surgery and length of stay for total knee arthroplasty. *Int J Med Inf* 158:104670. <https://doi.org/10.1016/j.ijmedinf.2021.104670>
- Han C, Liu J, Wu Y, Chong Y, Chai X, Weng X (2021) To predict the length of Hospital stay after total knee arthroplasty in an Orthopedic Center in China: the Use of Machine Learning algorithms. *Front Surg* 8:606038. <https://doi.org/10.3389/fsurg.2021.606038>
- Ramkumar PN, Navarro SM, Haeberle HS, Karnuta JM, Mont MA, Iannotti JP et al (2019) Development and validation of a machine learning Algorithm after primary total hip arthroplasty: applications to length of Stay and Payment models. *J Arthroplasty* 34:632–637. <https://doi.org/10.1016/j.arth.2018.12.030>
- Sridhar S, Whitaker B, Mouat-Hunter A, McCrory B (2022) Predicting length of Stay using machine learning for total joint replacements performed at a rural community hospital. *PLoS ONE* 17:e0277479. <https://doi.org/10.1371/journal.pone.0277479>
- Buddhiraju A, Chen TL-W, Subih MA, Seo HH, Esposito JG, Kwon Y-M (2023) Validation and generalizability of machine learning models for the prediction of Discharge Disposition following revision total knee arthroplasty. *J Arthroplasty* S0883–5403. <https://doi.org/10.1016/j.arth.2023.02.054>. 23)00185–7
- Chen TL-W, Buddhiraju A, Seo HH, Subih MA, Tuchinda P, Kwon Y-M (2023) Internal and External Validation of the Generalizability of Machine Learning Algorithms in Predicting Non-home Discharge Disposition following primary total knee Joint Arthroplasty. *J Arthroplasty* 38:1973–1981. <https://doi.org/10.1016/j.arth.2023.01.065>
- Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d'Alché-Buc F et al (2021) Improving reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 reproducibility program). *J Mach Learn Res* 22:1–20. <https://doi.org/10.48550/arXiv.2003.12206>
- Keswani A, Lovy AJ, Robinson J, Levy R, Chen D, Moucha CS (2016) Risk factors predict increased length of Stay and Readmission Rates in Revision Joint Arthroplasty. *J Arthroplasty* 31:603–608. <https://doi.org/10.1016/j.arth.2015.09.050>
- Chen TL-W, Buddhiraju A, Costales TG, Subih MA, Seo HH, Kwon Y-M (2023) Machine learning models based on a National-Scale Cohort identify patients at high risk for prolonged lengths of Stay following primary total hip arthroplasty. *J Arthroplasty* 38:1967–1972. <https://doi.org/10.1016/j.arth.2023.06.009>
- Chen TL-W, Wang Y, Peng Y, Zhang G, Hong TT-H, Zhang M (2023) Dynamic finite element analyses to compare the influences of customised total talar replacement and total ankle arthroplasty on foot biomechanics during gait. *J Orthop Transl* 38:32–43. <https://doi.org/10.1016/j.jot.2022.07.013>

28. Ferro CAT (2007) Comparing probabilistic forecasting systems with the Brier score. *Weather Forecast* 22:1076–1088. <https://doi.org/10.1175/WAF1034.1>
29. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak Int J Soc Med Decis Mak* 26:565–574. <https://doi.org/10.1177/0272989X06295361>
30. Cohen J (1988) *Statistical power analysis for the behavioral sciences*. 2 edition. Hillsdale, N.J: Routledge
31. Haeberle HS, Helm JM, Navarro SM, Karnuta JM, Schaffer JL, Callaghan JJ et al (2019) Artificial Intelligence and Machine Learning in Lower Extremity Arthroplasty: a review. *J Arthroplasty* 34:2201–2203. <https://doi.org/10.1016/j.arth.2019.05.055>
32. Klemm C, Tirumala V, Barghi A, Cohen-Levy WB, Robinson MG, Kwon Y-M (2022) Artificial intelligence algorithms accurately predict prolonged length of stay following revision total knee arthroplasty. *Knee Surg Sports Traumatol Arthrosc off J ESSKA* 30:2556–2564. <https://doi.org/10.1007/s00167-022-06894-8>
33. Garbarino LJ, Gold PA, Sodhi N, Anis HK, Ehiorobo JO, Boraiah S et al (2019) The effect of operative time on in-hospital length of stay in revision total knee arthroplasty. *Ann Transl Med* 7:66–66. <https://doi.org/10.21037/atm.2019.01.54>
34. Farley KX, Anastasio AT, Premkumar A, Boden SD, Gottschalk MB, Bradbury TL (2019) The influence of modifiable, postoperative patient variables on the length of stay after total hip arthroplasty. *J Arthroplasty* 34:901–906. <https://doi.org/10.1016/j.arth.2018.12.041>
35. Roger C, Debuyzer E, Dehl M, Bulaïd Y, Lamrani A, Havet E et al (2019) Factors associated with hospital stay length, discharge destination, and 30-day readmission rate after primary hip or knee arthroplasty: Retrospective Cohort Study. *Orthop Traumatol Surg Res OTSR* 105:949–955. <https://doi.org/10.1016/j.otsr.2019.04.012>
36. Bozic KJ, Kurtz SM, Lau E, Ong K, Chiu V, Vail TP et al (2010) The Epidemiology of Revision Total Knee Arthroplasty in the United States. *Clin Orthop* 468:45–51. <https://doi.org/10.1007/s11999-009-0945-0>
37. Matharu GS, Judge A, Murray DW, Pandit HG (2017) Outcomes following revision surgery performed for adverse reactions to metal debris in non-metal-on-metal hip arthroplasty patients: analysis of 185 revisions from the National Joint Registry for England and Wales. *Bone Jt Res* 6:405–413. <https://doi.org/10.1302/2046-3758.67.BJR-2017-0017.R2>
38. Grammatopoulos G, Pandit H, Kwon Y-M, Gundle R, McLardy-Smith P, Beard DJ et al (2009) Hip resurfacings revised for inflammatory pseudotumour have a poor outcome. *J Bone Joint Surg Br* 91:1019–1024. <https://doi.org/10.1302/0301-620X.91B8.22562>
39. Menken LG, Rodriguez JA (2020) Femoral revision for periprosthetic fracture in total hip arthroplasty. *J Clin Orthop Trauma* 11:16–21. <https://doi.org/10.1016/j.jcot.2019.12.003>
40. Chan VW, Chan P, Fu H, Cheung M, Cheung A, Yan C et al (2020) Preoperative optimization to prevent periprosthetic joint infection in at-risk patients. *J Orthop Surg* 28:230949902094720. <https://doi.org/10.1177/2309499020947207>
41. Rybak M, Lomaestro B, Rotschafer JC, Moellering R Jr, Craig W, Billeter M et al (2009) Therapeutic monitoring of Vancomycin in adult patients: a consensus review of the American Society of Health-System Pharmacists, the Infectious Diseases Society of America, and the Society of Infectious Diseases Pharmacists. *Am J Health Syst Pharm* 66:82–98. <https://doi.org/10.2146/ajhp080434>
42. Ling K, Tsouris N, Kim M, Smolev E, Komatsu DE, Wang ED (2023) Abnormal preoperative leukocyte counts and postoperative complications following total shoulder arthroplasty. *JSES Int* 7:601–606. <https://doi.org/10.1016/j.jseint.2023.03.001>
43. Kullo IJ, Hensrud DD, Allison TG (2002) Comparison of numbers of circulating blood monocytes in men grouped by body mass index (<25, 25 to <30, > or =30). *Am J Cardiol* 89:1441–1443. [https://doi.org/10.1016/s0002-9149\(02\)02366-4](https://doi.org/10.1016/s0002-9149(02)02366-4)

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.