Arthroplasty Today 10 (2021) 135–143

Contents lists available at ScienceDirect

Arthroplasty Today

journal homepage: http://www.arthroplastytoday.org/

Original research

A Novel, Potentially Universal Machine Learning Algorithm to Predict Complications in Total Knee Arthroplasty

Sai K. Devana, MD [a, *], Akash A. Shah, MD [a], Changhee Lee, MS [b], Andrew R. Roney, BA [a], Mihaela van der Schaar, PhD [b, c, d], Nelson F. SooHoo, MD [a]

[a] Department of Orthopaedic Surgery, University of California, Los Angeles, USA
[b] Department of Electrical and Computer Engineering, University of California, Los Angeles, USA
[c] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, London, UK
[d] The Alan Turing Institute, London, UK

ARTICLE INFO

Article history:
Received 24 March 2021
Received in revised form
23 June 2021
Accepted 25 June 2021
Available online xxx

Keywords:
AutoPrognosis
Machine learning
Predictive modeling
Knee replacement

ABSTRACT

Background: There remains a lack of accurate and validated outcome-prediction models in total knee arthroplasty (TKA). While machine learning (ML) is a powerful predictive tool, determining the proper algorithm to apply across diverse data sets is challenging. AutoPrognosis (AP) is a novel method that uses automated ML framework to incorporate the best performing stages of prognostic modeling into a single well-calibrated algorithm. We aimed to compare various ML methods to AP in predictive performance of complications after TKA.
Methods: Thirty-eight preoperative patient demographics and clinical features from all primary TKAs performed at California-licensed hospitals between 2015 and 2017 were evaluated as predictors of major complications after TKA. Traditional logistic regression (LR), various other ML methods (XGBoost, Gradient Boosting, AdaBoost, and Random Forest), and AP were used for model building to determine discriminative power (area under receiver operating curve), calibration (Brier score), and feature importance.
Results: Between 2015 and 2017, there were a total of 156,750 TKAs with 1109 (0.7%) total major complications. AP had the highest discriminative performance with area under receiver operating curve 0.679 compared with LR, XGBoost, Gradient Boosting, AdaBoost, and Random Forest (0.617, 0.601, 0.662, 0.657, and 0.545, respectively). AP (Brier score 0.007) had similar calibration as the other ML methods (0.006, 0.006, 0.022, 0.007, and 0.008, respectively). The variables that are most important for AP differ from those that are most important for LR.
Conclusion: Compared to conventional ML algorithms, AP has superior discriminative ability with similar calibration and suggests nonlinear relationships between variables in outcomes of TKA.

Introduction

Total knee arthroplasty (TKA) is a safe, cost-effective treatment for knee osteoarthritis that substantially improves quality of life and function in most patients [12,34]. Although the postoperative risk of mortality and major complications after elective TKA are generally low, substantial patient disability, dissatisfaction, and economic burden can result [6,10,21,22]. Recent prognostic research has focused on improving outcomes and patient satisfaction by guiding clinical decision with actionable predictive models of complications. Studies have identified patient and setting characteristics that are associated with higher risk [4,8,13,23,38,47], but few studies have developed and validated models that provide actionable intelligence regarding these risks in TKA. Accurate predictive modeling has the potential for improving preoperative decision-making, informed consent, and postoperative outcomes and can be essential for risk-adjustment of outcome-based performance measures and reimbursement programs that incentivize better performance on these outcome metrics [26].

Machine learning (ML) has shown promising results in generating predictive models that inform TKA treatment decisions and identify novel predictors of TKA outcomes better than traditional statistical methods [16,17,19,24,26,29,30,32,33]. However, existing

* Corresponding author. 10982 Roebling Ave, Apt 337, Los Angeles, CA 90024, USA. Tel.: +1-510-709-9494.
E-mail address: skdevana@gmail.com

https://doi.org/10.1016/j.artd.2021.06.020

models vary in performance according to the characteristics of each specific data set. Currently, application of ML involves manual selection of one or more ML algorithms which may not be the optimal choice for that particular data set.

AutoPrognosis (AP) is a novel algorithmic framework tailored for prognostic research that automatically selects and tunes the best possible ML algorithms and combines them into a single, well-calibrated predictive ensemble for any given data set [3]. Using a Bayesian optimization algorithm to efficiently configure the data set, AP combines the best-performing pipeline of ML algorithms into a single, well-calibrated predictive ensemble. This allows AP to be applied across a diverse group of data sets, circumventing the need for clinicians to choose the best algorithm. AP has shown promising results in other areas of medicine, outperforming other ML and traditional statistical modeling as well as uncovering novel predictive variables [1–3]. To the best of our knowledge, this is the first study to apply AP to TKA outcomes.

The aim of this study is to compare the performance of AP to that of traditional logistic regression (LR) and various other commonly used ML methods in predicting major complications after primary TKA. We hypothesized that AP would have the best predictive performance and be able to identify novel predictive variables.

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 allow patients to be tracked longitudinally for complications regardless of whether future admissions are at a different hospital from where the index procedure was performed.

Inclusion and exclusion criteria

The OSHPD database was used to retrospectively select patients older than 18 years from October 01, 2015, to December 13, 2017, that underwent elective primary TKA. Using International Classification of Diseases, Tenth revision, (ICD-10) Procedural codes, inclusion and exclusion criteria were based on the 2017 Procedure-Specific Measure Updates and Specifications Report Hospital-Level Risk-Standardized Complication Measure Version 6.0 measure developed by Centers for Medicare and Medicaid Services (CMS) International [49]. These criteria exclude patients with fracture of the pelvis or lower limbs coded in the principal or secondary discharge diagnosis fields of the index admission; a concurrent partial hip arthroplasty; a concurrent revision, resurfacing, or implanted device or prosthesis removal procedure; mechanical complications coded in the principal discharge diagnosis field; malignant neoplasm of the pelvis, sacrum, coccyx, lower limbs, or bone or bone marrow; or a disseminated malignant neoplasm coded in the principal discharge diagnosis field. All inclusion and exclusion ICD-10 codes are publicly available via CMS.

Outcome and other variables

The primary outcome measure was any major complication after index TKA (summarized in Table 1) which were also based on the CMS measures version 6.0 [49] and identified using the appropriate ICD-10-clincal modification codes. We excluded the measure for death within 30 days of the index admissions as death records were not available. The patient features and variables included in or derived from the OSHPD database are as follows: age, gender, race, ethnicity, income based on zip code of residence, teaching status of the hospital, rural location of the hospital, and total hospital volume of total joint arthroplasty (TJA) (TKA + total hip arthroplasty) from October 01, 2015, to December 13, 2017. Comorbidities were determined using the CMS-Condition Categories including metastatic cancer or acute leukemia, other major cancer, diabetes mellitus or diabetes mellitus complications, malnutrition, respiratory heart, digestive, urinary, other neoplasms, morbid obesity, bone, join, or muscle infections or necrosis, osteoarthritis (OA) of hip or knee (not associated with index procedure), rheumatoid arthritis and inflammatory connective tissue disease, osteoporosis and other bone or cartilage disorders, dementia or other specified brain disorders, major psychiatric disorders, hemiplegia/paraplegia/paralysis or functional disability, cardiorespiratory failure and shock, coronary atherosclerosis or angina, stroke, vascular or circulatory disease, chronic obstructive pulmonary disease, pneumonia, pleural effusions/pneumothorax, dialysis status, renal failure, decubitus ulcer or chronic skin ulcer, trauma, vertebral fractures without spinal cord injury, other injuries, skeletal deformities, and posttraumatic osteoarthritis (Table 4).

AutoPrognosis

AP is an algorithmic framework for automating the design of the ML-based clinical prognostic models [3] which eliminates the need for researchers and clinicians to have the in-depth knowledge of ML necessary to choose a particular prognostic model. A schematic illustration of AP is provided in Figure 1. AP uses Bayesian optimization techniques [46] to efficiently identify the ML pipelines (out of a huge space of possible pipelines; Table 2) that maximize a predefined performance metric. In this work, we use area under receiver operating curve (AUROC) performance, where a pipeline comprises design choices for classification methods and the corresponding hyperparameters. The Bayesian optimization algorithm used by AP implements a sequential exploration-exploitation scheme in which balance is achieved between exploring the utility (ie, AUROC) of new pipelines and re-examining the utility of previously explored ones. The motivation of using Bayesian optimization framework is due to its recent remarkable success in optimizing black-box functions with costly evaluations as compared to simpler approaches such as grid and random [46].

We implemented AP using the Python source code (https://bitbucket.org/mvdschaar/mlforhealthlabpub/src/68e4f7d13e4368eba655132a73ff9f278da5d3af/alg/autoprognosis/) of the original

Table 1
Complications.

Complications
Acute myocardial infarction: index admission or within 7 d of start of index admission
Pneumonia: index admission or within 7 d of start of index admission
Sepsis, septicemia, shock: index admission or within 7 d of start of index admission
Pulmonary embolism: index admission or within 30 d of start of index admission
Surgical site bleeding: index admission or within 30 d of index admission
Mechanical complications: index admission or within 90 d or start of index admission
Periprosthetic joint infection/wound infection: index admission or within 90 d of start of index admission

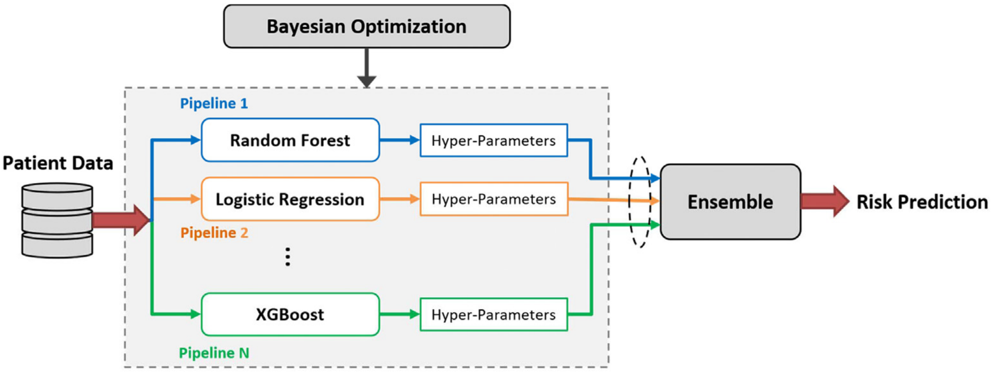

Figure 1. A schematic depiction of AutoPrognosis. AutoPrognosis is an automated framework that configures an optimally performing ensemble of ML-based prognostic models (various pipelines) to build a single well-calibrated algorithm for risk prediction.

paper. To train AP using Bayesian optimization, 5-fold stratified cross-validation on the training set (80% of the study population) was used to evaluate the performance of the pipeline (ie, ML pipeline of classification methods and the corresponding hyperparameters) in every iteration and to construct the ensemble model. Thus, the remaining held-out testing set (20% of the study population) is left unseen during training AP. We conducted 100 iterations of the Bayesian optimization procedure where the algorithm explores a new ML pipeline of classification methods and the corresponding hyperparameters in each iteration. Then, AP builds a final ensemble of all the ML pipelines that it explored through Bayesian optimization in which every pipeline is given a weight that is proportional to its empirical performance (Table 3).

Statistical analysis

The primary outcome measure was any major complication after index TKA. We evaluated the discriminative and calibration performances of the models under consideration via 5-fold stratified cross-validation on the overall cohort. In every cross-validation fold, the training cohort (80% of the study population) was used to derive our model (AP) and the ML benchmark models, and then a held-out testing cohort (20% of the study population) was used for performance evaluation.

We considered 5 standard ML benchmarks that cover different classes of ML modeling approaches to compare against AP as follows: LR (linear classifier), random forest [11] (a tree-based ensemble classifier), AdaBoost [40], gradient boosting machines [31] (Gradient Boosting), and XGBoost [14] (boosting ensemble classifiers). The purpose of including these models individually in our analysis is to show AP automatically selects and tunes the best possible model which outperforms these individually tuned ML models. We implemented LR, Random Forest, AdaBoost, and Gradient Boosting machines using the *scikit-learn* Python library [39] and XGBoost using the *xgboost* Python library [14]. The hyperparameters 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 and 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.

For AP and the 5 aforementioned ML models, discrimination (which assesses how well a model distinguishes patients who developed postoperative complications and those who did not) was assessed using 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. A value greater than 0.9 is considered to have high discriminative power, 0.7-0.9 indicates moderate discriminative power, and 0.5-0.7 indicates low discriminative power [18]. Calibration, which assesses the agreement between predictions and the observed outcomes (postoperative complications), was assessed using Brier scores and corresponding calibration plots. Brier score provides 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 indicate better calibration of the prognostic model. We also performed post-hoc discriminative power analysis (AUROC) on subgroup cohorts of patients with diabetes and obesity to assess for improved performance given these groups had higher percentage of patients with at least 1 complication than the overall cohort.

We used the partial dependence function introduced in the article by Friedman et al. in 2001 [31] to measure the importance of

Table 2
List of classification methods in AutoPrognosis.

Classification methods		
Logistic regression	Random forest	Gradient boosting
eXtreme Gradient Boosting (XGBoost)	AdaBoost	Bagging
Bernoulli Naïve Bayes	Gaussian Naïve Bayes	Multinomial Naïve Bayes
Perceptron	Decision Trees	Support Vector Machine (SVM)
Latent Dirichlet Allocation (LDA)	Quadratic Discriminant Analysis (QDA)	K-Nearest Neighbors (KNN)
Neural Networks		

Table 3
List of the 10 pipelines fitted to the TKA cohort.

Pipeline #	Methods	Hyperparameters	Weight
1	Random Forest	(max_depth = 5, n_estimators = 98)	0.199
2	Random Forest	(max_depth = 5, n_estimators = 96)	0.191
3	Random Forest	(max_depth = 5, n_estimators = 102)	0.170
4	Random Forest	(max_depth = 3, n_estimators = 101)	0.155
5	Logistic Regression	(l2-penalty, 0.139)	0.091
6	Logistic Regression	(l2-penalty, 0.231)	0.075
7	AdaBoost	(n_estimators = 150)	0.063
8	XGBoost	(max_depth = 5, n_estimators = 153)	0.045
9	Logistic Regression	(l2-penalty, 0.029)	0.007
10	Gradient Boosting	(max_depth = 5, n_estimators = 96)	0.005

an individual feature or variable 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.

AUROC and Brier scores were reported as mean values with standard deviations and 95% confidence intervals for all models. Feature importance was reported as numerical values.

Results

Demographic characteristics

Between October 01, 2015, to December 13, 2017, there was a total of 156,750 elective primary TKAs, the majority of which were females (61.4%). Patient age ranged from 18 to 100 years with a median age of 68 years. There were a total of 1109 (0.7%) complications (989 patients had at least 1 complication), with pneumonia being the most common complication. The overall demographics, patient features, and complications are summarized in Table 4.

Table 4
Patient demographics and overall complications.

Variable	All patients (n = 156,750)
Age range of patients (y)	18-100
Mean age ± SD (y)	68.2 ± 9.2
Median age (y)	68
Males	60464 (38.6%)
Females	96286 (61.4%)
Race	
Black	8764 (5.6%)
Native American	580 (0.4%)
Asian or Pacific Islander	8832 (5.6%)
White	116954 (74.6%)
Other	18260 (11.6)
Unknown	3360 (2.1%)
Ethnicity	
Hispanic	29480 (18.8%)
Non-Hispanic	125521 (80.1%)
Unknown	1749 (1.1%)
Hospital volume range[a]	1-8149
Mean hospital volume ± SD[a]	1854.6 ± 1568
Insurance	
Medicare	93461 (59.6%)
Medical	10264 (6.5%)
Workers compensation	5398 (3.4%)
Other	2016 (1.3%)
Private	45771 (29.2%)
Comorbidities (CMS Clinical Conditions)	
Metastatic cancer	164 (0.1%)
Other major cancer	1920 (1.2%)
Neoplasms	1250 (0.8%)
Diabetes mellitus	32991 (21%)
Malnutrition	672 (0.4%)
Morbid obesity	16818 (10.7%)
Rheumatoid arthritis	6713 (4.2%)
Osteoarthritis	2626 (1.7%)
Osteoporosis	14226 (9.1%)
Dementia	1529 (1%)
Major psychiatric disorder	7537 (4.8%)
Paralysis	232 (0.1%)
Coronary artery disease or dangina	13668 (8.7%)
COPD	7890 (5%)
Renal failure	12469 (7.9%)
Decubitus ulcer	89 (5.7 × 10^{-2}%)
Vertebral fracture	36 (2.3 × 10^{-2}%)
Skeletal deformities	27 (1.7 × 10^{-2}%)
Posttraumatic OA	76 (4.8 × 10^{-2}%)
Total complications	1109 (0.7%)
# of Patients having at least 1 complication	989 (0.6%)
AMI	91 (5.8 × 10^{-2}%)
Pneumonia	474 (0.3%)
Sepsis	201 (0.1%)
PE	273 (0.2%)
Surgical site bleeding	13 (6.3 × 10^{-4}%)
Mechanical complications	31 (2.0 × 10^{-2}%)
Infection	26 (1.7 × 10^{-2}%)

AMI, acute myocardial infarction; COPD, chronic obstructive pulmonary disease; PE, pulmonary embolism; SD, standard deviations.

[a] Hospital volume is the total number of TJA cases performed between October 01, 2015, to December 13, 2017.

Model performance and calibration

In predicting patients having at least one postoperative complication, AP had the highest discriminative performance with AUROC 0.68 ± 0.04 compared with LR, XGBoost, Gradient Boosting, AdaBoost, and Random Forest (0.63 ± 0.01, 0.60 ± 0.03, 0.66 ± 0.04, 0.66 ± 0.03, 0.55 ± 0.02, respectively; Table 5). The AUROC performance gain of AP was statistically significant compared with LR ($P < .05$), XGBoost ($P < .01$), and Random Forest ($P < .001$). The gain over Gradient Boosting and AdaBoost was not statistically significant. In regard to calibration performance, AP (Brier score 0.0067 ± 0.0010) was similar to the other ML methods (0.0063 ± 0, 0.0065 ± 0.0020, 0.0072 ± 0.0031, 0.0072 ± 0, 0.0075 ± 0.0002, respectively; Table 5). Figure 2 outlines the similarity in calibration plots, specifically of AP and LR as an example.

The diabetes cohort (n = 32,991) and obesity cohort (n = 16,818) had higher percentage of patients with at least 1 complication (0.8% and 1%, respectively) than the overall cohort (0.6%). AP performed better in the overall cohort (0.679 ± 0.04) than the obesity (0.660 ± 0.02) and diabetes (0.657 ± 0.04) subgroups (Table 5). However, gain in AUROC from the best performing ML model (Gradient Boosting) relative to AP was higher in the obesity (0.026) and diabetes (0.02) cohorts than that in the overall cohort (0.017) despite being significantly smaller populations.

Feature Importance

The relative importance of each variable (binary, categorical, and continuous) to the model performance for AP and LR is displayed in Figure 3. The variables that are most important for AP differ from those that are most important for LR.

Discussion

Owing to increasing demand and excellent outcomes, the annual number of primary TKAs is projected to grow 85% to 1.26 million procedures by 2030 [45]. This further amplifies the substantial cost and morbidity caused by the inevitable associated increase in postoperative complications and unplanned readmissions. Accurate statistical prediction tools are thus valuable in improving preoperative counseling, informed consent, shared decision-making, postoperative expectations, and risk-adjusted reimbursement programs.

Substantial effort has gone into developing various prediction models of outcomes in orthopedic surgery [9,15,25,41]. The American College of Surgeons National Surgical Quality Improvement Program developed a universal surgical risk calculator using an extensive database across multiple specialties (only 12% of which were orthopedic procedures) which has shown to have good overall accuracy averaged across procedures [5]. However, performance studies of this model for orthopedic procedures, including TJA, are limited to single-site cohorts showing fair to poor results [15,48].

While TJA-specific preoperative risk prediction models have been developed, they have shown poor or unknown performance on internal or external validation [35]. The American Joint Replacement Registry Risk Calculator which estimates risk for 90-day mortality and 2-year prosthetic joint infection was developed

Table 5
Discriminative power and calibration.

Methods	AUROC overall (n = 156,750)	Brier score overall (n = 156,750)	AUROC obesity (n = 16,818)	AUROC diabetes (n = 32,991)
Logistic Regression	0.629 ± 0.01 (0.604-0.654)	0.006 ± 0 (0.0063-0.0063)	0.619 ± 0.03 (0.602-0.636)	0.583 ± 0.07 (0.526-0.640)
XGBoost	0.601 ± 0.03 (0.578-0.624)	0.006 ± 0.0002 (0.0063-0.0066)	0.567 ± 0.03 (0.540-0.594)	0.590 ± 0.05 (0.549-0.630)
Gradient Boosting	0.662 ± 0.04 (0.625-0.698)	0.022 ± 0.0031 (0.0051-0.0106)	0.634 ± 0.04 (0.601-0.666)	0.637 ± 0.05 (0.594-0.680)
AdaBoost	0.657 ± 0.03 (0.630-0.684)	0.007 ± 0 (0.0072-0.0072)	0.625 ± 0.02 (0.609-0.641)	0.635 ± 0.03 (0.605-0.665)
Random Forest	0.545 ± 0.02 (0.525-0.565)	0.008 ± 0.0002 (0.0073-0.0077)	0.534 ± 0.03 (0.508-0.559)	0.549 ± 0.05 (0.505-0.593)
AutoPrognosis	0.679 ± 0.04 (0.642-0.716)	0.007 ± 0.0010 (0.0058-0.0075)	0.660 ± 0.02 (0.646-0.674)	0.657 ± 0.04 (0.620-0.693)

All values reported as mean ± standard deviation with (95% confidence interval).

without reporting accuracy metrics [8]. This model was externally validated using a sample of Medicare eligible patients from Veterans Health Administration and found to have very poor accuracy (0.62 C-statistic) for 90-day mortality [25]. Recently, Harris et al. reported fair-to-moderate accuracy for prediction of 30-day mortality and cardiac complications (C statistic 0.73 and 0.75, respectively), whereas prediction of deep vein thrombosis and reoperation was poor (C statistic 0.59 and 0.6, respectively) after TKA and total hip arthroplasty with internal and external validation of their model [24]. While these results are promising, acceptable prediction accuracy was limited only to a subset of the complications studied.

The aforementioned risk-prediction tools were developed with various individual ML learning methods (LR, boosted regression, least absolute shrinkage, and selection operator). Here we report the use of a novel ensemble ML algorithm for predicting complications after TKA using the OSHPD database containing over 150,000 patients. Compared to the aforementioned existing TJA ML models in the literature which have shown fair to moderate performances (AUROC > 0.7), our AP model in this study had poor performance (between 0.5 and 0.7) likely due to the low-quality billing-based data set used. However, with our particular data set, AP demonstrated superior discriminative power relative to traditional LR and 4 other standard ML algorithms. Differential performance relative to LR of the other 4 ML models highlights the importance of correct ML method selection and appropriate adjusting of hyperparameters [3] for a particular data set. With similar calibration and superior discrimination compared with the other ML methods, AP demonstrated superior prediction performance for patients having at least one complication after TKA in the OSHPD data set. The complication-prediction calculator built based off the AP model in our study is available at https://risk-calculator-tka-comp.herokuapp.com. Users simply plug in available patient data to obtain percent risk of a major complication.

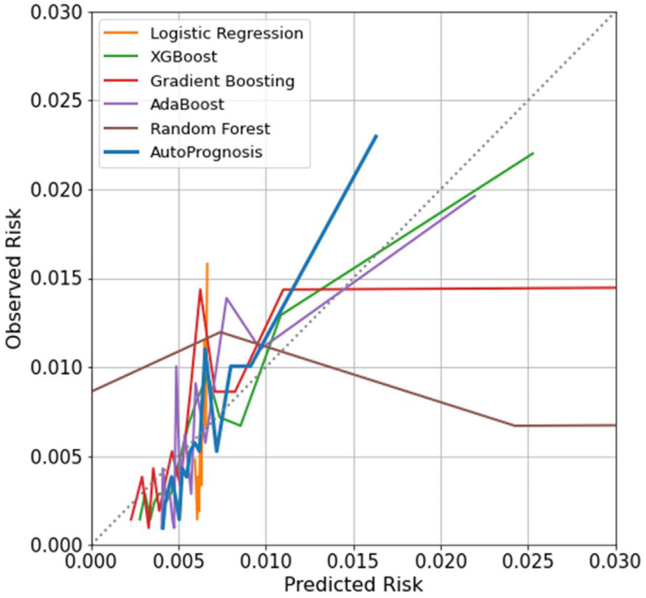

Figure 2. Calibration plot. Calibration, measure of how close the predicted risk is to the observed risk, is similar between AutoPrognosis and Logistic Regression.

We also report the relative importance of several features to performance of the AP algorithm. 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. Of the binary features and variables, malnutrition was the most important. Preoperative hypoalbuminemia (<3.5 grams/deciliter) is an accepted marker of malnutrition and has been widely shown to be a strong risk factor for postoperative mortality, morbidity, readmission, and increased length of stay specifically in primary TJA [7,20]. Decubitus ulcer had the next highest feature importance which, to our knowledge, is not widely reported in the literature as a robust independent risk factor for postoperative complications after TKA. While the presence of decubitus ulcer may indicate a lack of baseline mobility and function and lead most surgeons to not recommend TKA on a particular patient to begin with, it is still important to note that this (presence of decubitus ulcer preoperatively) is an important feature in those patients that have undergone TKA. We identified dementia as the third most important feature for this algorithm which similar to malnutrition has been previously shown to have an association with increased postoperative complications and resource utilization [8,28]. In regard to continuous variables, hospital volume was more important with both AP and LR than age. Increasing age and lower hospital volume have both been shown to be associated with higher risk of surgical site infections, complications, and mortality [8,37,42,44]. We must note that AP is currently unable to generate threshold predictive values above which the risk for a complication is significantly increased. This is because the model does not treat continuous variables (ie, age, hospital volume) as independent risk factors but rather generates a risk of complication based on the values of all the explanatory features. Finally, there was a difference in feature importance of categorical variables of insurance status, race, and ethnicity when comparing LR and AP. While there is some overlap of feature importance from AP vs LR, the 2 models differ in their relative importance of features (Fig. 3). This highlights the value of AP as it is able to elucidate complex nonlinear relationships (depending on the combination and relative weighting of pipelines used).

In addition to the overall cohort, we performed a post-hoc subgroup analysis on the diabetes and obesity cohorts to assess AP performance on smaller populations with higher complication rates. These cohorts were chosen because of their increasing prevalence in the population and their well-known association

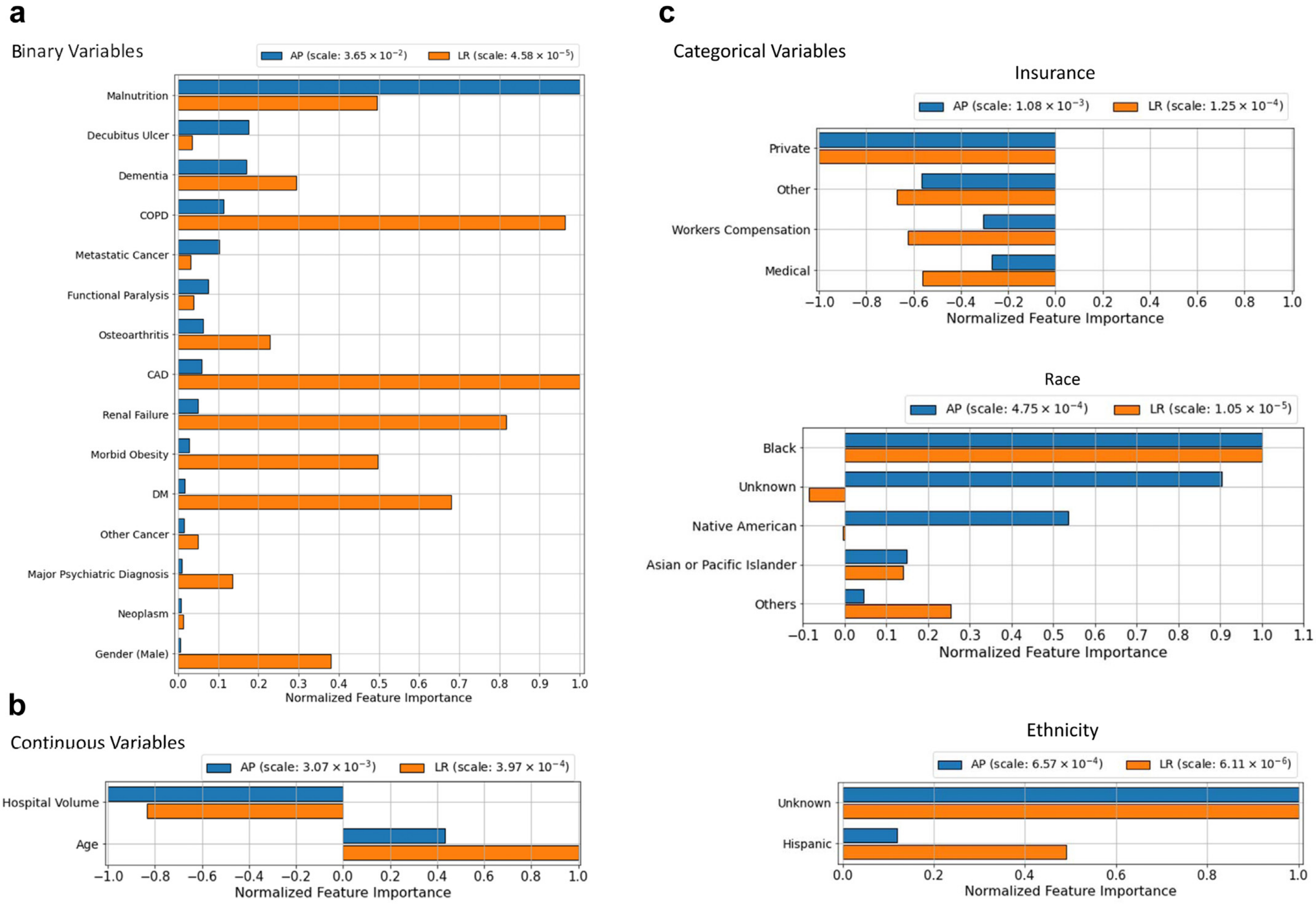

Figure 3. Feature importance. The 15 most important binary features to AutoPrognosis are shown with their respective LR feature importance (a). Autoprognosis and logistic regression have differing feature importance for binary (a) variables as well as continuous (b) and categorical (c) variables. This suggests important nonlinear relationships that are captured by AutoPrognosis that Logistic Regression cannot. COPD, chronic obstructive pulmonary disease; DM, diabetes mellitus; CAD, coronary artery disease.

with higher risk of postoperative adverse outcomes [36,43]. Interestingly, obesity and diabetes were the 10th and 11th most important features (respectively) in terms of predictive performance on the AP algorithm as compared to fourth and sixth most important features for the LR algorithm (Fig. 3a). The fact that these features were relatively less important for AP again demonstrates that advanced ensemble models may better identify important feature interactions. While AUROC was lower for the diabetes and obesity subgroups, the gain (difference in AUROC) between AP and the next best performing ML method, Gradient Boosting, was higher in the 2 subgroups than that in the overall cohort. Thus, AP has an improvement in performance and could be more useful in populations with a higher incidence of the outcome of interest.

The retrospective nature of this study inherently lends itself to limitations. Although OSHPD has a large patient sample, patient data (such as body mass index, hemoglobin A1C, smoking status, orthopedic complications), surgeon factors (experience, fellowship-trained, and so forth), and outcomes collected are limited. The use of ICD diagnosis 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 orthopedic complications, mortality, patient-reported functional outcomes, and patient satisfaction. Owing 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 data set 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 5 prognostic models in terms of area under precision-recall curve (which can be a more sensitive performance metric in the imbalanced setting), which again showed relatively superior performance of AP (Appendix II). Along the same lines, owing to the low overall complication rate, secondary analysis of individual complications and outcomes was beyond the scope of our model (although we certainly recognize the clinical importance of such results). It is important to note that surgical outcome prediction models built on retrospective data only include those patients that have presumably undergone preoperative risk stratification by their surgeon and primary care physician introducing a selection bias from the general population requesting a consultation for surgical evaluation. While the primary aim of this study was to show superior performance of AP relative to other ML models, the overall performance of AP against the existing ML models in the literature is poor and has low clinical utility. This could be due to the low granularity of patient features inherent to billing databases as well as the low rate of complications. We believe that AP would perform better in a data set with more outcomes of interest as the data would be more balanced which would prevent overfitting and increase heterogeneity (helping the ML model to find differences between patients). An imbalanced data set can lead to overfitting and, in this case, overcalling noncomplication as the model is used for 99.4% noncomplications. It should be noted that any predictive algorithm is only as reliable as the data it is built on. Furthermore, systemic biases in clinical decisions and data collection are amplified by ML, potentially adversely affecting historically underrepresented groups such as patients of lower socioeconomic status, ethnic minorities, and women [27]. Future studies should thus validate or refute the models in this analysis by using more granular multi-institutional data or prospective evaluation. The authors wish to reiterate that this is a proof of principle study aimed at evaluating the efficacy of AP compared with traditional ML methods. Finally, we must acknowledge the black box nature of ML algorithms that can lead to nonphysiologic patient features that effect a very small portion of the cohort (ie, malnutrition and decubitus ulcer) having high significance, which highlights that ML provides predictive modeling at the expense of statistical inference of clinical outcomes.

Conclusions

Here we report the use of a novel ensemble ML algorithm for prediction of major complications after primary TKA. AP is unique in its utility of an automated ML framework to incorporate the best performing stages of existing ML algorithms into a single well-calibrated algorithm. Using AP, we developed an algorithm that was well calibrated while showing superior discrimination compared with other individual ML methods. While the AP model in this study was modest in its predictive performance and may not dramatically change clinical practice, continuing to apply this modeling technique to diverse data sets for multiple outcomes can have promising clinical implications. Ultimately, development of accurate predictive modeling is helpful for preoperative counseling, informed consent, shared decision-making, risk adjustment reimbursement programs, and guiding surgeon and patient postoperative expectations. AP is a versatile tool that can be used to identify important patient features in predicting outcomes across diverse data sets to ultimately improve patient outcomes.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Acknowledgments

This research was in part supported by the H. H. Lee Research Grant, Department of Orthopedics, UCLA.

References

[1] Alaa AM, Bolton T, Angelantonio E Di, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PLoS One 2019;14(5):e0213653.
[2] Alaa AM, van der Schaar M. Prognostication and risk factors for cystic fibrosis via automated machine learning. Sci Rep 2018;8(1):11242.
[3] Alaa AM, Van Der Schaar M. Autoprognosis: automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. In: 35th international conference on machine learning. Stockholmsmässan, Stockholm SWEDEN: ICML; 2018.
[4] Belmont PJ, Goodman GP, Waterman BR, Bader JO, Schoenfeld AJ. Thirty-day postoperative complications and mortality following total knee arthroplasty: incidence and risk factors among a national sample of 15,321 patients. J Bone Joint Surg Am 2014;96(1):20.
[5] Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg 2013;217(5):833.
[6] Blom AW, Brown J, Taylor AH, Pattison G, Whitehouse S, Bannister GC. Infection after total knee arthroplasty. J Bone Joint Surg Br 2004.
[7] 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.
[8] Bozic KJ, Lau E, Kurtz S, et al. Patient-related risk factors for periprosthetic joint infection and postoperative mortality following total hip arthroplasty in medicare patients. J Bone Joint Surg Am 2012;94(9):794.
[9] 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.
[10] Bozic KJ, Ries MD. The impact of infection after total hip arthroplasty on hospital and surgeon resource utilization. J Bone Joint Surg Am 2005.
[11] Breiman L. Random forests. Mach Learn 2001;45:5.
[12] Canovas F, Dagneaux L. Quality of life after total knee arthroplasty. Orthop Traumatol Surg Res 2018;104(1):S41.
[13] Chamieh JS, Tamim HM, Masrouha KZ, Saghieh SS, Al-Taki MM. The association of anemia and its severity with cardiac outcomes and mortality after total knee arthroplasty in noncardiac patients. J Arthroplasty 2016;31(4):766.
[14] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, CA: ACM Digital Library; 2016.
[15] 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.
[16] Ehlers AP, Roy SB, Khor S, et al. Improved risk prediction following surgery using machine learning algorithms. EGEMS 2017;5(2):3.
[17] Farooq H, Deckard ER, Ziemba-Davis M, Madsen A, Meneghini RM. Predictors of patient satisfaction following primary total knee arthroplasty: results from a traditional statistical model and a machine learning algorithm. J Arthroplasty 2020:30618.
[18] Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. Intensive Care Med 2003.
[19] Fontana MA, Lyman S, Sarker GK, Padgett DE, MacLean CH. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? Clin Orthop Relat Res 2019;477(6):1267.
[20] Fryhofer GW, Sloan M, Sheth NP. Hypoalbuminemia remains an independent predictor of complications following total joint arthroplasty. J Orthop 2019;16(6):552.
[21] Gandhi R, Davey JR, Mahomed NN. Predicting patient dissatisfaction following joint replacement surgery. J Rheumatol 2008;35(12):2415.
[22] Harmelink KEM, Zeegers AVCM, Hullegie W, Hoogeboom TJ, Nijhuis-van der Sanden MWG, Staal JB. Are there prognostic factors for one-year outcome after total knee arthroplasty? A systematic review. J Arthroplasty 2017;32(12):3840.
[23] Harris A, Reeder R, Ellerbe L, Bradley K, Rubinsky A, Giori N. Preoperative alcohol screening scores. J Bone Jt Surg 2011;93(4):321.
[24] Harris AH, Kuo AC, Bowe T, Gupta S, Nordin D, Giori NJ. Prediction models for 30-day mortality and complications after total knee and hip arthroplasties for veteran health administration patients with osteoarthritis. J Arthroplasty 2018;33(5):1539.
[25] 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.
[26] Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori NJ. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? Clin Orthop Relat Res 2019;477(2):452.
[27] Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery. Ann Surg 2018;268:70.
[28] Hernandez NM, Cunningham DJ, Jiranek WA, Bolognesi MP, Seyler TM. Total knee arthroplasty in patients with dementia. J Knee Surg 2020;24(4):265.
[29] 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.
[30] 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.
[31] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29(5):1189.
[32] Kunze KN, Karhade AV, Sadauskas AJ, Schwab JH, Levine BR. Development of machine learning algorithms to predict clinically meaningful improvement for the patient-reported health state after total hip arthroplasty. J Arthroplasty 2020;35(8):2119.
[33] Liang ZC, Wang W, Murphy D, Hoi J, Hui P, Surgery J. TI novel coronavirus and orthopaedic surgery AR TI. J Bone Joint Surg Am 2020;1:1.
[34] Losina E, Walensky RP, Kessler CL, et al. Cost-effectiveness of total knee arthroplasty in the United States: patient risk and hospital volume. Arch Intern Med 2009;169(12):1113.
[35] Manning DW, Edelstein AI, Alvi HM. Risk prediction tools for hip and knee arthroplasty. J Am Acad Orthop Surg 2016;24(1):19.
[36] Martin JR, Jennings JM, Dennis DA. Morbid obesity and total knee arthroplasty: a growing problem. J Am Acad Orthop Surg 2017;25(3):188.

[37] Murphy BPS, Dowsey MM, Choong PFM. The impact of advanced age on the outcomes of primary total hip and knee arthroplasty for osteoarthritis: a systematic review. JBJS Rev 2018;6(2):e6.

[38] Parvizi J, Sullivan TA, Trousdale RT, Lewallen DG. Thirty-day mortality after total knee arthroplasty. J Bone Joint Surg Am 2001;83(8):1157.

[39] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(1):2825.

[40] Rätsch G, Onoda T, Müller KR. Soft margins for AdaBoost. Mach Learn 2001;42(3):287.

[41] 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.

[42] Santaguida PL, Hawker GA, Hudak PL, et al. Patient characteristics affecting the prognosis of total hip and knee joint arthroplasty: a systematic review. Can J Surg 2008;51(6):428.

[43] Shohat N, Goswami K, Tarabichi M, Sterbis E, Tan TL, Parvizi J. All patients should Be screened for diabetes before total joint arthroplasty. J Arthroplasty 2018;33(7):2057.

[44] Singh JA, Kwoh CK, Boudreau RM, Lee GC, Ibrahim SA. Hospital volume and surgical outcomes after elective hip/knee arthroplasty: a risk-adjusted analysis of a large regional database. Arthritis Rheum 2011;63(8):2531.

[45] Sloan M, Premkumar A, Sheth NP. Projected volume of primary total joint arthroplasty in the u.s., 2014 to 2030. J Bone Joint Surg Am 2018;100(17):1455.

[46] Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. Adv Neural Inf Process Syst 2012:2951.

[47] Weaver F, Hynes D, Hopkinson W, et al. Preoperative risks and outcomes of hip and knee arthroplasty in the veterans health administration. J Arthroplasty 2003;18(6):693.

[48] Wingert NC, Gotoff J, Parrilla E, Gotoff R, Hou L, Ghanem E. The ACS NSQIP risk calculator is a fair predictor of acute periprosthetic joint infection. Clin Orthop Relat Res 2016;474(7):1643.

[49] Yale New HAven Heatlh Services/Center for Outcomes Research & Evaluation (YNHHSC/CORE). Procedure-specific measure Updates and Specifications report hospital-level risk-standardized complication measure. 2017. https://qualitynet.cms.gov/inpatient/measures/complication/methodology. [Accessed 25 July 2021].

Appendix I

Supplementary Table 1
AUPRC performance (mean and 95% CI).

Models	Aurpc
Logistic Regression	0.015 (0.011-0.018)
XGBoost	0.013 (0.010-0.015)
Gradient Boosting	0.020 (0.013-0.027)
AdaBoost	0.022 (0.016-0.028)
Random Forest	0.008 (0.007-0.009)
AutoPrognosis	0.025 (0.018-0.032)

We used the partial dependence function introduced in the study by Friedman et al. in 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 $\mathscr{X}$ and $\mathscr{X}_{\backslash c}$ be its complement, ie, $\mathscr{X} = \mathscr{X}_{\backslash c} \cup x_c$, and $r(\mathscr{X}) = r(\mathscr{X}_{\backslash c}, x_c)$ be the predicted risk by our trained model. Then, we define the feature importance score for an individual feature x_c by averaging $r(\mathscr{X}_{\backslash c}, x_c = 1) - r(\mathscr{X}_{\backslash c}, x_c = 0)$ for binary features and $r(\mathscr{X}_{\backslash c}, x_c = \max(x_c)) - r(\mathscr{X}_{\backslash c}, x_c = \min(x_c))$, 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 $r(\mathscr{X}_{\backslash c}, x_c = b) - r(\mathscr{X}_{\backslash c}, x_c = \text{mode}(x_c))$, where $\text{mode}(x_c)$ indicates the most frequency category of feature x_c.

Appendix II

Given the low complication rate of our cohort, our data may be imbalanced which can be a limitation. However, we believe that models trained with an artificially balanced data set (via oversampling samples with complications or via downsampling samples with no complications) cannot be directly used in the clinical setting. Although balancing the data set may improve the discriminative power of trained models, it distorts the true data distribution. Thus, the models trained on the balanced data set will provide poor calibration on the held-out testing set, which is assumed to have the same distribution as that of the unseen patients once the models are deployed.

Supplementary Table 2
Confusion matrix for AutoPrognosis.

Auto prognosis	True condition	
Prediction condition	True positive 177	False positive 22,591
	False negative 21	True negative 8561

We acknowledge that area under receiver operating curve performance may not be the best performance metric for comparing different prognostic models in an imbalanced setting. To address this concern, we evaluated the prognostic models considered in this study in terms of the area under precision-recall curve; Supplementary Table 1. This can be a more sensitive performance metric when comparing different models within an imbalanced data set.

Although the AUPRC performance of AutoPrognosis and that of other prognostic models considered in this study are all poor, AutoPrognosis provides relatively superior performance when compared to the baseline (ie, random guessing based on the observed frequency of having complications) which is 0.006. This implies that if we set the threshold value for converting continuous predictions into binary predictions relatively low (down to the observed frequency), we can sacrifice the precision of the model (due to the increased false positives) and instead improve the sensitivity. For example, when the threshold value is set to 0.006, AutoPrognosis achieved the sensitivity score of 0.894 while the logistic regression model achieved the sensitivity score of 833; please see the confusion matrices (Supplementary Tables 2 and 3) for more details.

Supplementary Table 3
Confusion matrix for logistic regression.

Logistic regression	True condition	
Prediction condition	True positive 165	False positive 23,357
	False negative 33	True negative 7795