ORIGINAL CLINICAL RESEARCH REPORT

OPEN

Machine Learning-Based Models Predicting Outpatient Surgery End Time and Recovery Room Discharge at an Ambulatory Surgery Center

Rodney A. Gabriel, MD, MAS,*†‡ Bhavya Harjai, BS,‡ Sierra Simpson, PhD,§ Nicole Goldhaber, MD,‖ Brian P. Curran, MD,* and Ruth S. Waterman, MD, MS*

BACKGROUND: Days before surgery, add-ons may be scheduled to fill unused surgical block time at an outpatient surgery center. At times, outpatient surgery centers have time limitations for end of block time and discharge from the postanesthesia care unit (PACU). The objective of our study was to develop machine learning models that predicted the following composite outcome: (1) surgery finished by end of operating room block time and (2) patient was discharged by end of recovery room nursing shift. We compared various machine learning models to logistic regression. By evaluating various performance metrics, including F1 scores, we hypothesized that models using ensemble learning will be superior to logistic regression.

METHODS: Data were collected from patients at an ambulatory surgery center. The primary outcome measurement was determined to have a value of 1 (versus 0) if they met both criteria: (1) surgery ends by 5 PM and (2) patient is discharged from the recovery room by 7 PM. We developed models to determine if a procedure would meet both criteria if it were scheduled at 1 PM, 2 PM, 3 PM, or 4 PM. We implemented regression, random forest, balanced random forest, balanced bagging, neural network, and support vector classifier, and included the following features: surgery, surgeon, service line, American Society of Anesthesiologists score, age, sex, weight, and scheduled case duration. We evaluated model performance with Synthetic Minority Oversampling Technique (SMOTE). We compared the following performance metrics: F1 score, area under the receiver operating characteristic curve (AUC), specificity, sensitivity, precision, recall, and Matthews correlation coefficient.

RESULTS: Among 13,447 surgical procedures, the median total perioperative time (actual case duration and PACU length stay) was 165 minutes. When SMOTE was not used, when predicting whether surgery will end by 5 PM and patient will be discharged by 7 PM, the average F1 scores were best with random forest, balanced bagging, and balanced random forest classifiers. When SMOTE was used, these models had improved F1 scores compared to no SMOTE. The balanced bagging classifier performed best with F1 score of 0.78, 0.80, 0.82, and 0.82 when predicting our outcome if cases were to start at 1 PM, 2 PM, 3 PM, or 4 PM, respectively.

CONCLUSIONS: We demonstrated improvement in predicting the outcome at a range of start times when using ensemble learning versus regression techniques. Machine learning may be adapted by operating room management to allow for a better determination whether an add-on case at an outpatient surgery center could be appropriately booked. (Anesth Analg 2022;135:159–69)

KEY POINTS

- **Question:** By using machine learning, can we more accurately predict whether a surgical add-on for an outpatient surgery center would both end at the predetermined block time end and the patient would be discharged from the recovery room at a predetermined time point?
- **Findings:** We developed a predictive model using Synthetic Minority Oversampling Technique (SMOTE) and balanced bagging techniques that improved the ability to predict the timing outcome at a range of start times allowing for better scheduling.
- **Meaning:** Enhanced modeling and prediction methods will improve patient care, staff scheduling, and institutional profits.

GLOSSARY

ASA PS = American Society of Anesthesiologists Physical Status; **AUC** = area under the receiver operating characteristic curve; **EQUATOR** = Enhancing the Quality and Transparency of Health Research; **MCC** = Matthews correlation coefficient; **OR** = operating room; **PACU** = postanesthesia care unit; **ROC** = receiver operating characteristic; **SD** = standard deviation; **SMOTE** = Synthetic Minority Oversampling Technique

Efficiency is paramount to the success of an ambulatory surgery center, since it is a major financial contributor to many health care organizations.[1–3] Important metrics used to quantify efficiency include first-case delays,[4] turnover times,[5] case duration booking accuracy,[6,7] and postanesthesia care unit (PACU) length of stay.[8,9] Various studies have developed predictive models for improving accuracy of case duration booking,[6,7,10,11] as such strategies may help improve operating room (OR) utilization.

While most studies focus on case duration[6,7] and PACU length of stay,[8,9] more analyses are needed to develop models that can predict both case duration and PACU length of stay for a given patient. This is especially true in an outpatient surgery center, which has hard deadlines for surgical blocks and PACU nursing staff hours before incurring overages. PACU length of stay should also be considered given that there may also be a specific time when recovery room staffing is scheduled to end. Schedulers must also allow for time so that the patient can safely recover and be discharged within the correct time window.

During the operational stage of perioperative management, short-term managerial decisions—such as scheduling cases, staffing ORs, and focusing efforts on finishing the day on schedule, are often made far in advance of the surgery date.[12] However, it is not always possible to perfectly assign enough cases within appropriate blocks. For outpatient surgeries, OR managers are often faced with filling unused time with only days before date of surgery. Booking surgeries during this time should put into account both predicted case duration as well as PACU length of stay to meet the time constraints set at each unique institution. While there have been several predictive models estimating case duration,[6,7] there have been none described that can predict both: (1) if the final case in an OR will finish by a specified time (ie, end of surgical block time) and (2) if the patient will be discharged from the PACU by the end of the day (ie, time at which outpatient surgery closes).

The utilization of machine learning to aid in OR management has much potential, as various studies have demonstrated improvement in case duration accuracy, cancelation prevention, and recovery room management.[13] The objective of our study is to assess whether we can develop predictive models that can help managers determine if an add-on would finish by the end of the operating day and have the patient discharged on time. The primary outcome is defined as, given a specified start time, whether the surgery will finish by the end of the scheduled block time (ie, 5 PM) and patient will be discharged from the recovery room when staffing is scheduled to finish (ie, 7 PM). Here, we use ensemble learning combined with an oversampling technique and compare models using various performance metrics, including the F1 score, which is the harmonic mean of precision and recall. Whether more advanced machine learning algorithms can outperform logistic regression for clinical prediction in patient datasets is uncertain; however, the benefit may vary based on the type of data, outcomes studied, and patient population.[14] We thus hypothesized that machine learning approaches using ensemble learning—such as random forest, balanced bagging, or balanced random forests—would outperform models using logistic regression (which may only model linear relationships between dependent and independent variables).

METHODS

Study Sample

This retrospective study was approved by the Human Research Protections Program at the University of California San Diego for the collection of data from our electronic medical record system. For this study, the informed consent requirement was waived. Data from all patients who underwent outpatient surgery from our institution's freestanding ambulatory surgery center from March 2018 to January 2021 were extracted retrospectively. Cases that occurred <10 times were removed from the analysis. The manuscript adheres to the applicable Enhancing the Quality and Transparency of Health Research (EQUATOR) guidelines for observational studies.

Primary Objective and Data Collection

The primary outcome measurement was binary and was determined to have a value of 1 if they met both criteria: (1) surgery ended by 5 PM and (2) patient was discharged from the PACU by 7 PM. These times were chosen because this was our institutional times

From the *Department of Anesthesiology, University of California, San Diego, La Jolla, California; †Division of Biomedical Informatics, Department of Medicine, University of California, San Diego, La Jolla, California; ‡Division of Perioperative Informatics, Department of Anesthesiology, University of California, San Diego, La Jolla, California; §Department of Psychiatry, University of California, San Diego, La Jolla, California; and ‖Department of Surgery, University of California, San Diego, La Jolla, California.

Accepted for publication February 16, 2022.

Funding: Support was provided solely from institutional and/or departmental sources.

Conflicts of Interest: See Disclosures at the end of the article.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (www.anesthesia-analgesia.org).

Reprints will not be available form the authors.

Address correspondence to Rodney A. Gabriel, MD, MAS, University of California, San Diego, 9400 Campus Point Dr, La Jolla, CA 92037. Address e-mail to ragabriel@health.ucsd.edu.

DOI: 10.1213/ANE.0000000000006015

for end of surgical block time and when our recovery room team was expected to be done (before overtime pay was instituted). We developed predictive models to determine if a specific case would meet both criteria if they were scheduled at 1 PM, 2 PM, 3 PM, or 4 PM (Figure 1A). We implemented logistic regression, random forest classifier, support vector classifier, simple-feedforward neural network, balanced random forest classifier, and balanced bagging classifier. In addition, we evaluated model performance with and without Synthetic Minority Oversampling Technique (SMOTE) (Figure 1B). Model performance was primarily evaluated by quantifying the F1 score, which is the harmonic mean of precision and recall (defined in detail below).

The independent features included in the models were surgical procedure, surgeon identification, American Society of Anesthesiologists Physical Status (ASA PS) score, age (years), sex, weight (kg), surgical service line, scheduled surgical incision time, and scheduled room time (time for patient to enter OR to time they leave OR). In addition, we collected actual room time and actual PACU length of stay. Of note, we did not include anesthesiologist identification nor primary anesthesia type as these features may generally not be known days before surgery (we included only the features that can be defined days beforehand). The primary outcome was defined as positive if the actual room time did not extend past 5 PM (based on the start

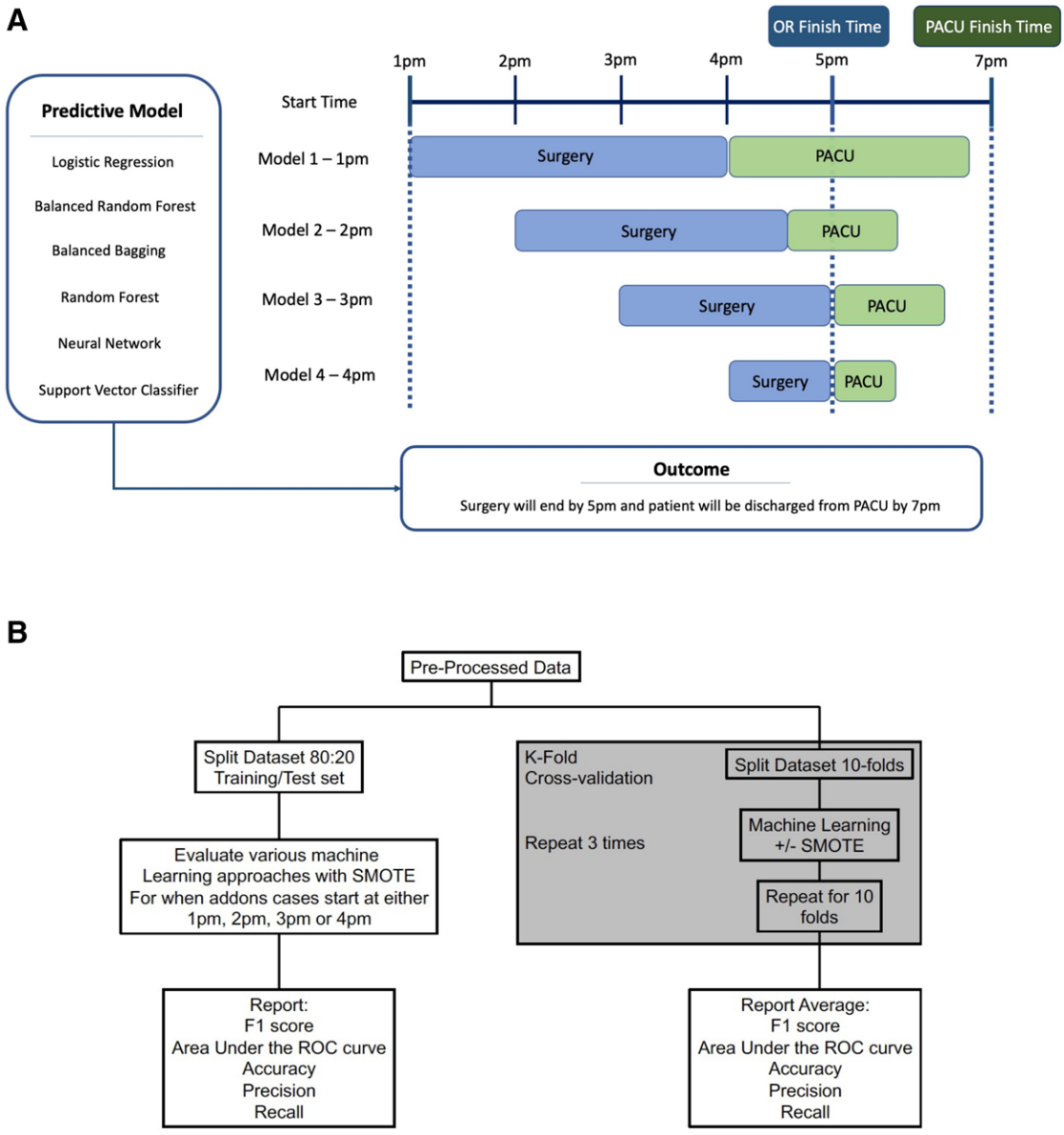

Figure 1. Overview of study methodology. A, An illustration of the study design describing the use of machine learning to predict the outcome (defined as surgery ending by 5 PM and patient discharged from recovery room by 7 PM). Each model was based on whether the start time of the surgery was 1 PM, 2 PM, 3 PM, or 4 PM. B, Data pipeline. OR indicates operating room; PACU, postanesthesia care unit; ROC, receiver operating characteristic; SMOTE, synthetic minority oversampling technique.

time) and the patient was discharged from the PACU by 7 PM. To calculate whether patient was discharged by 7 PM, we calculated the total perioperative time, which was defined as total minutes from start time to actual discharge (based on both actual case duration and PACU length of stay). For example, if a surgery was to start at 1 PM, then the total perioperative time must be <6 hours (out of PACU by 7 PM), and the case duration needs to be <4 hours (out of OR by 5 PM). There were no missing data. All the data captured were consistently documented in the electronic medical record system; which included surgeon name, surgical procedure, service line, scheduled surgery time, actual surgery time, PACU length of stay, patient ASA class, age, sex, and weight. Therefore, no imputation was performed.

Statistical Analysis

Python (v3.7.5) was used for all statistical analyses. First, the cohort was divided into training and test data sets, reflecting an 80:20 split using a randomized splitter—the "train_test_split" method from the scikit learn library[15]—thus, any patient present in the test set was automatically removed from the training set. We developed each machine learning model using the training set (using SMOTE) and tested its performance on the test set (measuring F1 score, recall, precision, Matthews correlation coefficient, sensitivity, specificity, and area under the receiver operating characteristic curve [AUC] for receiver operating characteristic [ROC] curve). Second, we then calculated the average F1 score, recall, precision, Matthews correlation coefficient, sensitivity, specificity, and AUC using k-folds cross-validation (described below).

Data Balancing. SMOTE for Nominal and Continuous algorithm—implemented using the "imblearn" library—was used to create a balanced class distribution.[16] Imbalanced data may be particularly difficult for predictive modeling due to the uneven classification of data. A balanced dataset would have minimal difference in positive and negative outcomes. However, if the difference is large, it is considered unbalanced. SMOTE is a statistical technique that increases the number of cases in the minority dataset to balance it with the majority dataset—while not affecting the number of majority cases. This algorithm takes samples of the feature space for each target class and 5 of its nearest neighbors and then generates new cases that combine features of the target case with features of its nearest neighbors. This method increases the percentage of the minority cases in the dataset and allows for improved downstream analysis. Crucially, SMOTE was applied only to our training sets, and we did not oversample the testing set, thus maintaining the natural outcome frequency.

Machine Learning Models. We evaluated 6 different classification models: logistic regression, random forest classifier, support vector classifier, simple-feed-forward neural network, balanced random forest classifier, and balanced bagging classifier. For each, we also compared the use of oversampling the training set via SMOTE versus no SMOTE. For each model, all features were included as inputs. For each machine learning model, we performed hyperparameter tuning via grid search (described below for each model) before performing the final version on that model. Feature importance was ranked based on Gini importance.

Multivariable Logistic Regression. This is a statistical model that asserts a binary outcome based on the weighted combination of the underlying independent variables. We tested an L2-penalty-based regression model without specifying individual class weights. This model provided a baseline score and helped make the case for improvement over the evaluation metrics. For hyperparameter tuning, we performed grid search cross-validation to find the optimal parameter value for C (the inverse of regularization strength), which was 0.1.

Random Forest Classifier. We developed a random forest classifier, and the criterion for the split was set to the Gini impurity. The Gini impurity was calculated in which C is the total number of classes, and $p(i)$ was the probability of picking a datapoint with class i.

$$C\ G = i = \sum p(i) * (1 - p(i)) \qquad 1$$

Random forest is an ensemble approach (a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model) of decision trees, which by themselves have been proven to work well for a variety of classification problems.[17] The random forest is a robust and reliable nonparametric supervised learning algorithm that acts as a means to test the further improvement in the metrics and provide the feature importance of the dataset. For hyperparameter tuning, we performed grid search cross-validation to find the optimal parameter value for maximum depth, minimal samples required to be at the leaf node, minimal samples required to split an internal node, and the number of estimates (ie, number of trees), in which the values were 50, 2, 5, and 500, respectively.

*Support Vector Classifier*A support vector classifier tries to find a hyperplane decision boundary that best splits the data into the required number of classes. It plots each data item as a point in an n-dimensional space (n being the number of features), then finds a

hyperplane that separates the classes.[18] We developed a modification of the support vector machine that weighed the margin proportional to the class importance and was a cost-sensitive support vector classifier (by choosing the gamma value as "scale," while defining the classifier and assigning "balanced" to the class-weight parameter).

(gamma: defines the amount of influence a single data point can have, "scale": assigns gamma value as 1/[*n*_features * X.var()])

For hyperparameter tuning, we performed grid search cross-validation to find the optimal parameter value for *C* (value that trades off correct classification of training examples against maximation of the decision function's margin), which was 10.

Balanced Random Forest Classifier. This is an implementation of the random forest, which randomly undersamples each bootstrap to balance it. The model was built using the imblearn package. For hyperparameter tuning, we performed grid search cross-validation to find the optimal parameter value for maximum depth, minimal samples required to be at the leaf node, minimal samples required to split an internal node, and the number of estimates (ie, the number of trees), in which the values were 50, 2, 5, and 500, respectively.

Balanced Bagging Classifier. Another way to ensemble models is bagging or bootstrap-aggregating. Bagging methods build several estimators on different randomly selected subsets of data. Unlike random forests, bagging models are not sensitive to the specific data on which they are trained. They give the same score even when trained on a subset of the data. Bagging classifiers are also generally more immune to overfitting. We built a balanced bagging classifier using the imblearn package. For hyperparameter tuning, we performed grid search cross-validation to find the optimal parameter values for maximum samples and the number of estimators, which were 0.05 and 500, respectively.

Multilayer Perceptron Neural Network. Using sci-kit learn's "MLPClassifier," we built a basic shallow feed-forward network. The activation function was set to the rectified linear unit function, and the net was trained for a maximum of number iterations based on hyperparameter tuning. For hyperparameter tuning, we performed grid search cross-validation to find the optimal parameter values for the number of hidden layers, the number of neurons per hidden layer, and the maximum number of iterations, which were 2, 10, and 500, respectively.

PERFORMANCE METRICS

Our primary performance metric of interest was the F1 score. This is a version of the Fβ-metric, where we provided equal weight to the precision and recall scores. F1 score is formally equal to the harmonic mean of precision and recall, and this provides a way to combine both into a single metric. This is the most valuable metric to analyze a classification task and, thus, was the most significant metric of our analysis.[19,20]

$$\text{F1 Score} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$$

In addition, we reported precision, recall, AUC for the ROC curve, the Matthews correlation coefficient, sensitivity, and specificity.[21]

K-Folds Cross-Validation. To perform a more robust evaluation of our models, we implemented repeated stratified-K-fold cross-validation to observe the precision, recall, F1 score, and AUC scores for 10 splits and 3 repeats. For each iteration, the dataset was split into 10 folds, where 1 fold served as the test set, and the remaining 9 sets served as the training set. The model was built on the training set. In the case when SMOTE was utilized, only the training set was oversampled. This was repeated until all folds had the opportunity to serve as the test set. This was then repeated 3 times. For each iteration, our performance metrics were calculated on the test set. The average of each performance metric was calculated thereafter.

RESULTS

There were a total of 13,447 surgical procedures included in our analysis, which mostly comprised orthopedic (36%) and ear, nose, and throat (13.1%) surgeries. The overall median (quartiles) OR case time was 74.5 (49–113) minutes, and PACU length of stay was 84 (63–112) minutes. The overall median (quartiles) total perioperative time (actual case duration and PACU length stay) was 165 (122–225) minutes (Table 1).

First, we split the data into training and test sets and calculated the AUC of the ROC curve for each machine learning approach on the test set using SMOTE. We built a model for each start time (1 PM, 2 PM, 3 PM, or 4 PM) and showed that the ensemble learning approaches (ie, balanced bagging classifier, balanced random forest, and random forest) had the highest AUC scores (Figure 2). For example, if a case is booked at 2 PM and we would like to determine if it would finish before 5 PM and patient discharged by 7 PM, the logistic regression model had an AUC of 0.81, while balanced bagging classifier was 0.91. Based on the balanced random forest classifier, the most important features (ranked by the Gini importance index) contributing to the models were scheduled room times, scheduled incision time, surgical specialty, and patient weight (Figure 3). As an example,

when predicting the outcome with a 3 PM start, the association of each covariate to the outcome based on the logistic regression was presented in Supplemental Digital Content 1, Table 1, http://links.lww.com/AA/D916.

Performance Metrics Calculated From K-Folds Cross-Validation

We then calculated the average F1 score, accuracy, precision, recall, and AUC from k-folds cross-validation among all models with or without SMOTE (Table 2). We calculated the F1 scores of each model with and without SMOTE for each start time point (Figure 4, dotted lines). When SMOTE was not utilized, the random forest classifier and balanced bagging classifier F1 scores performed the best among all start time points. For example, at 3 PM start time, the F1 scores were 0.69, 0.65, and 0.69 for logistic regression, multilayer perceptron, and support vector classifier, respectively, while the F1 scores were 0.74, 0.77, and 0.80 for balanced random forest, balanced bagging, and random forest classifiers, respectively.

When SMOTE was utilized, we noted some improvements in F1 scores for balanced bagging classifier, balanced random forest, and random forest. Furthermore, these models had the highest F1 scores at all start time points (Figure 4, solid lines). For example, for 3 PM start time, the F1 scores were 0.67, 0.69, and 0.70 for logistic regression, multilayer perceptron, and support vector classifier, respectively, while the F1 scores were 0.80, 0.82, and 0.81 for balanced

Table 1. Distribution of Data. This Includes All Features Included in the Model Except for Actual Surgical Procedure and Surgeon Identification

	All cases	
Characteristic	**n**	**%**
Total	13,447	
Service line		
Other	1183	8.8
Breast surgery	866	6.4
Colorectal surgery	1203	8.9
Ears nose and throat	1766	13.1
Minimally invasive surgery	745	5.5
Obstetrics and gynecology	1904	14.2
Orthopedic surgery	4843	36.0
Urology	937	7.0
ASA PS ≥ 3 (%)	2895	21.5
Age (y)–mean (SD)	49.4 (16.6)	
Male sex (%)	5409	40.2
Weight (kg)–mean (SD)	79.0 (19.2)	
Number of cases in the operating room–median (quartiles)	5 (4–7)	
Number of times surgeon performed surgery–median (quartiles)	38 (11–96)	
Scheduled incision time (min)–median (quartiles)	60 (33–90)	
Scheduled room time (min)–median (quartiles)	65 (40–95)	
Actual room time (min)–median (quartiles)	74.5 (49–113)	
PACU length of stay (min)–median (quartiles)	84 (63–112)	
Total perioperative time (room + PACU min)–median (quartiles)	165 (122–225)	

Abbreviations: ASA PS, American Society of Anesthesiologists Physical Status; PACU, postanesthesia care unit; SD, standard deviation.

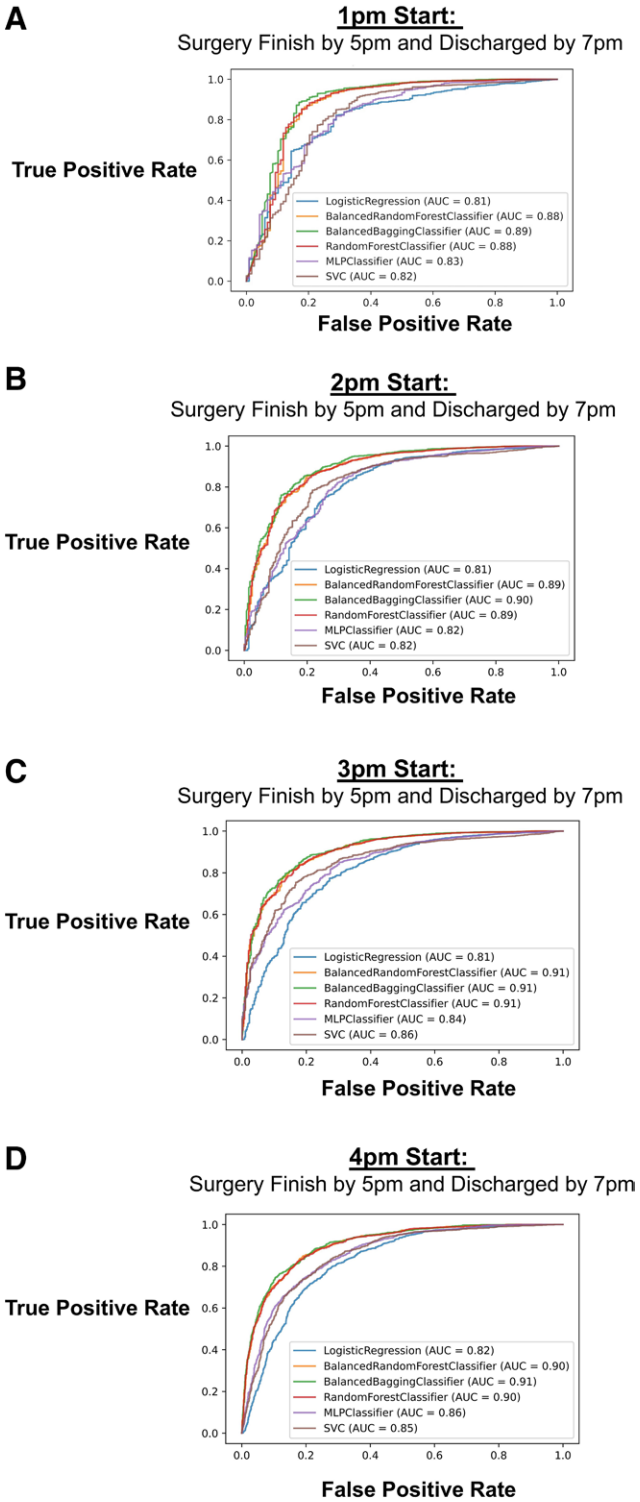

Figure 2. AUC for 6 separate models: logistic regression, multilayer perceptron neural network classifier, balanced random forest, balanced bagging classifier, random forest classifier, and support vector classifier. The models were implemented to predict the outcome for when a procedure will start at: A, 1 pm; B, 2 pm; C, 3 pm; or D, 4 pm. AUC indicates area under the receiver operating characteristic curve; SVC, support vector classifier.

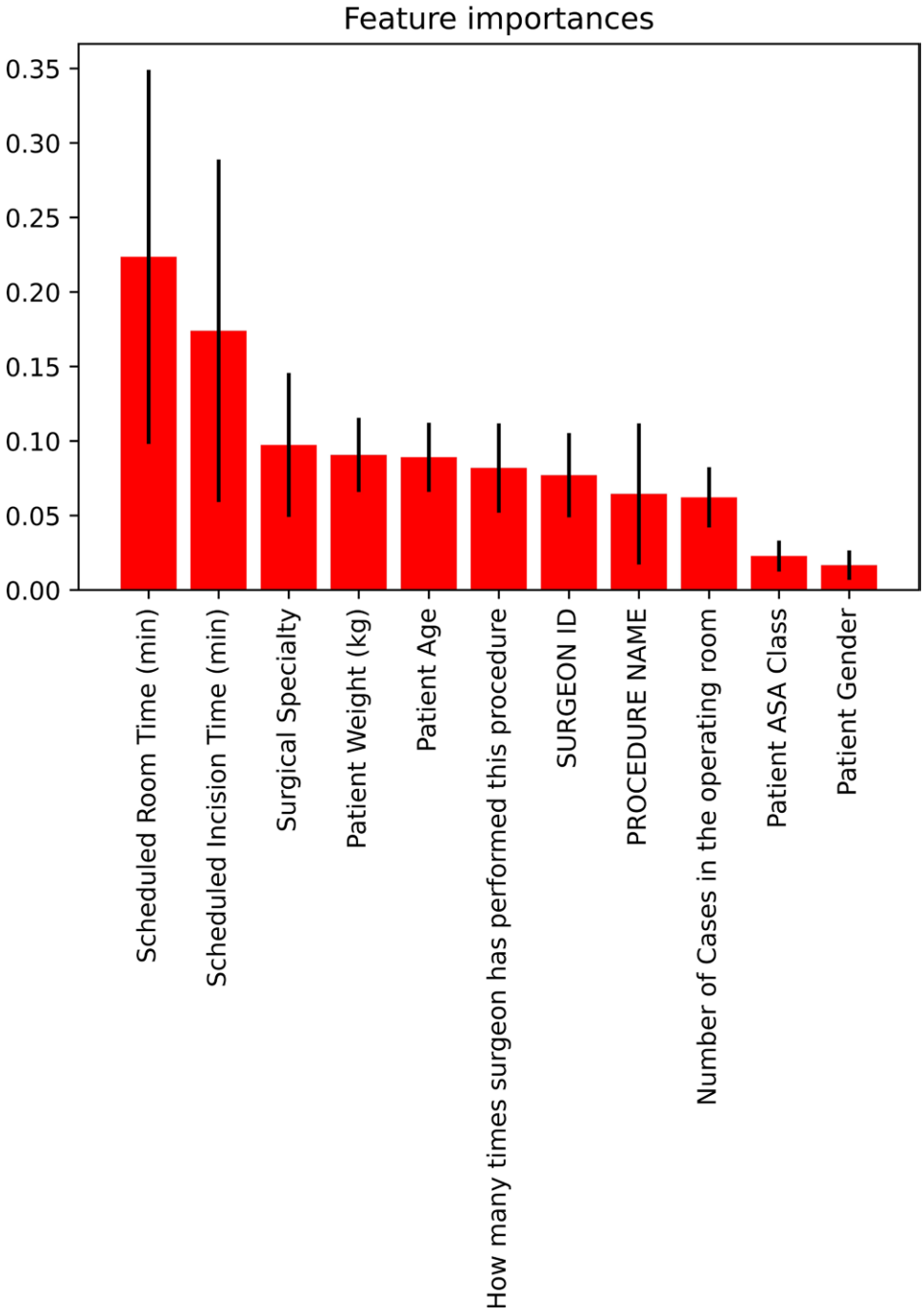

Figure 3. Feature importance graph of 11 features based on the balanced bagging approach. ASA indicates American Society of Anesthesiologists.

random forest, balanced bagging, and random forest classifiers, respectively. The balanced bagging classifier performed best with F1 score of 0.78, 0.80, 0.82, and 0.82 when predicting our outcome if cases were to start at 1 PM, 2 PM, 3 PM, or 4 PM, respectively. This was compared to F1 score of 0.56, 0.62, 0.67, and 0.73 for logistic regression.

DISCUSSION

In our analysis, we compared various types of machine learning approaches to predict, given a specific start time, if surgery will finish before end of surgical block time and if the patient will also be discharged by the end of the recovery room team shift. We found that prediction is improved using an oversampling technique, which balances training data sets of surgeries of asymmetrical incidence. We demonstrated improvement in predicting the outcome at a range of start times, which would allow for greater flexibility in scheduling.

In many ambulatory outpatient surgery practices, cases are booked into a specified surgical block

Table 2. Average Performance Metrics (Precision, Recall, MCC, Cohen's Kappa, Sensitivity, Specificity, and AUC) for Each Machine Learning Modeling Predicting Surgery Ending by 5 PM and Patient Discharged From PACU by 7 PM Based on 1 PM, 2 PM, 3 PM, or 4 PM Start

	Without SMOTE						With SMOTE					
Classifier	Precision	Recall	MCC	Sensitivity	Specificity	AUC	Precision	Recall	MCC	Sensitivity	Specificity	AUC
1 PM start: predicting if surgery will end before 5 PM and patient discharged from recovery room before 7 PM												
Logistic regression	0.956	0.996	0.152	0.996	0.064	0.809	0.986	0.784	0.269	0.784	0.766	0.829
Balanced random forest classifier	0.991	0.855	0.378	0.855	0.832	0.913	0.979	0.974	0.516	0.984	0.564	0.919
Balanced bagging classifier	0.991	0.882	0.414	0.883	0.824	0.919	0.981	0.979	0.555	0.978	0.583	0.928
Random forest classifier	0.968	0.996	0.480	0.996	0.309	0.929	0.979	0.974	0.517	0.974	0.569	0.919
Multilayer perceptron neural network	0.959	0.994	0.231	0.996	0.123	0.844	0.984	0.838	0.295	0.838	0.711	0.847
Support vector classifier	0.956	0.998	0.149	0.998	0.049	0.724	0.844	0.849	0.327	0.848	0.751	0.858
2 PM start: predicting if surgery will end before 5 PM and patient discharged from recovery room before 7 PM												
Logistic regression	0.909	0.982	0.327	0.982	0.222	0.798	0.950	0.766	0.317	0.766	0.685	0.796
Balanced random forest classifier	0.972	0.832	0.476	0.832	0.811	0.899	0.952	0.952	0.576	0.952	0.625	0.903
Balanced bagging classifier	0.969	0.870	0.512	0.870	0.782	0.905	0.953	0.961	0.604	0.961	0.624	0.912
Random forest classifier	0.933	0.982	0.543	0.982	0.445	0.909	0.952	0.951	0.572	0.951	0.622	0.901
Multilayer perceptron neural network	0.912	0.980	0.352	0.980	0.253	0.824	0.953	0.800	0.358	0.800	0.691	0.828
Support vector classifier	0.914	0.981	0.381	0.981	0.275	0.769	0.961	0.814	0.405	0.814	0.739	0.840
3 PM start: predicting if surgery will end before 5 PM and patient discharged from recovery room before 7 PM												
Logistic regression	0.888	0.970	0.433	0.970	0.353	0.822	0.935	0.780	0.394	0.780	0.713	0.816
Balanced random forest classifier	0.963	0.815	0.524	0.815	0.835	0.910	0.937	0.941	0.609	0.941	0.661	0.910
Balanced bagging classifier	0.961	0.853	0.563	0.853	0.816	0.916	0.936	0.952	0.632	0.952	0.656	0.919
Random forest classifier	0.918	0.975	0.608	0.975	0.539	0.918	0.937	0.941	0.611	0.941	0.664	0.910
Multilayer perceptron neural network	0.891	0.971	0.448	0.971	0.365	0.847	0.938	0.805	0.427	0.805	0.716	0.846
Support vector classifier	0.886	0.980	0.447	0.980	0.329	0.809	0.952	0.783	0.453	0.783	0.789	0.855
4 PM start: predicting if surgery will end before 5 PM and patient discharged from recovery room before 7 PM												
Logistic regression	0.674	0.588	0.449	0.588	0.846	0.821	0.736	0.764	0.465	0.764	0.720	0.809
Balanced random forest classifier	0.701	0.833	0.620	0.833	0.808	0.900	0.829	0.775	0.632	0.775	0.861	0.901
Balanced bagging classifier	0.726	0.817	0.634	0.817	0.832	0.903	0.837	0.773	0.642	0.773	0.871	0.905
Random forest classifier	0.779	0.732	0.629	0.732	0.887	0.902	0.829	0.774	0.629	0.774	0.860	0.901
Multilayer perceptron neural network	0.699	0.640	0.501	0.640	0.849	0.843	0.771	0.794	0.523	0.794	0.749	0.849
Support vector classifier	0.708	0.670	0.527	0.670	0.850	0.844	0.767	0.817	0.534	0.817	0.740	0.846

Abbreviations: AUC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient; PACU, postanesthesia care unit; SMOTE, synthetic minority oversampling technique.

time to allow for appropriate surgical duration and recovery time. However, the entire block is not always filled due to asymmetrical service time. At a prespecified time point, the surgical block is often opened to fill the block to optimal utilization. Thus, add-ons may be requested days before surgery. When

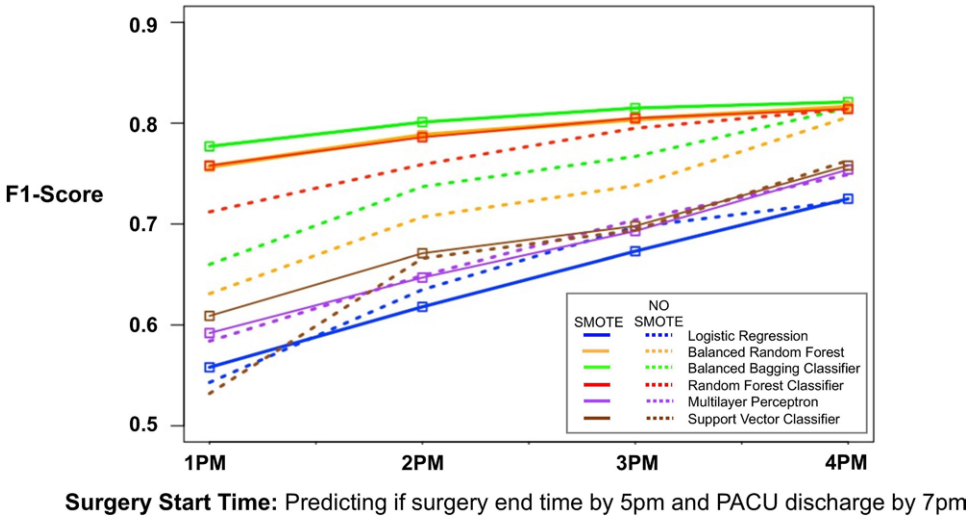

Figure 4. The F1 score (calculated by k-folds cross-validation) for each machine learning approach predicting whether booking a surgical procedure at 1 PM, 2 PM, 3 PM, or 4 PM will lead to surgery ending by 5 PM and patient discharged from recovery room by 7 PM (solid lines, SMOTE; dotted lines, no SMOTE). PACU indicates post-anesthesia care unit; SMOTE, synthetic minority oversampling technique.

scheduling add-ons, it is important to keep in mind the nursing and anesthesiology staffing restrictions. Overextending staff is costly and inefficient. Using the type of analysis proposed in this study, OR managers can then better decide where specific surgery add-ons should be placed or not placed.

Advances in electronic health monitoring have resulted in data that are better curated and easier to access. This opportunity lends itself to advance analysis of big data with artificial intelligence approaches to improve outcomes. Limited resources can then be directed to the most impactful changes to optimize efficiency. Using a single dataset, we can improve our ability to predict outcomes by using machine learning approaches to determine the most efficacious approach to scheduling. Such practice can be implemented as standard practice in the health care setting.

The primary objective of our study was to compare various machine learning models to a base model using logistic regression. In our study, ensemble learning (random forest, balanced random forest, and balanced bagging classifier) approaches performed better than logistic regression in predicting total perioperative time (case duration and PACU length of stay), when comparing F1 scores. At all starting time points, the F1 score was higher for random forest, balanced random forest, and balanced bagging compared to logistic regression. Furthermore, when SMOTE was used, the improvement in F1 scores for this model was more apparent. A major limitation of the binomial logistic regression approach used in this study is that this type of regression assumes linearity between the dependent and independent variables.[22,23] Ensemble learning models are advantageous in that they can leverage multiple learning algorithms to better predict an outcome and can also capture nonlinear relationships between features and outcomes that may contribute to the improved performance. Ensemble learning approaches are not without limitations—they rely on diversity within the sample and between the models; however, methods to introduce diversity can be used to generate diversity and class balance within a given dataset. Based on our results, we recommend the use of ensemble learning approaches versus logistic regression for tackling this type of operating problem. More studies are needed to assess the feasibility of implementing machine learning approaches to improve outpatient surgery efficiency; however, to be successful, features included in the model need to be data that are easily automatically collected via the electronic medical record system (such as the case with the features we chose in our analysis).

Additionally, a common issue with electronic health data is class imbalance.[24] Expansion of the minority class can be achieved by synthetic generation of minority class data by SMOTE, which uses a nearest-neighbor approach to reduce the class imbalance. Standard methods replicate random data from the original dataset, which can induce bias from a small collection of individuals. In the present study, SMOTE improved our model performance when directly compared to regression techniques. This is likely because an SMOTE balanced dataset better represents the initial population than other approaches.

It is important to note that the most important predictors that were included in the models in this study were scheduled room times, scheduled incision time, surgical specialty, and patient weight. Surgical specialty played a role, as different surgical specialties had more consistently timed or "standard" procedures and/or patient populations compared to other specialties. Patient weight was a significant contributor as it can take added time/resources to physically move the patient, perform the surgery, or establish an airway.[25]

There are metrics, including OR block time utilization and scheduled case duration accuracy, which are used to evaluate an institution's efficiency. However, a lack of standardization among institutions makes comparison difficult. In a previous single-center study at this institution,[26] it was determined using "time stamp data" taken directly from the electronic medical records that scheduled case duration accuracy was the most positively associated metric with scheduled end-time accuracy. Similarly, it has been demonstrated in a multicenter study that underfilled OR block time contributed more to OR utilization and efficiency compared to turnover time or on-time case start.[27] Many strategies have been used to optimize the filling of OR block time, including changing the sequence of cases in order of duration,[28,29] staffing similar cases with similar staff,[30] or using historical data from specific surgeons[31–34] to improve case duration accuracy. Data mining with predictive modeling[34–36] has been trialed as well. Attempts have been made to use machine learning and artificial intelligence to improve the accuracy case duration estimates,[6,7,37–40] but no one has used artificial intelligence to predict the feasibility of end of day case add-ons combined with PACU length of stay.

There are some limitations to consider from this study. As a retrospective study, collection and accuracy of the data are only as reliable as it was collected. This is single institution data; therefore, our data will not capture interinstitution differences. Furthermore, there are no universal time points for end of surgical block time nor end of PACU staffing; thus, the times we used for this analysis (5 PM for end of surgical block time and 7 PM for end of PACU staffing) may not apply widely. However, the focus of the study was to set an example of how one can use predictive modeling to accurately estimate whether surgeries will end by institution-specific block time end, and patients will be discharged by institution-specific end of PACU staffing time. Nonetheless, we are focusing our novel findings based on the improvement of prediction using our methodological approach of combining ensemble learning and SMOTE. SMOTE is an effective method for limiting class imbalance; however, it can be problematic for high-dimensional data.

Health care organizations rely on ambulatory surgery centers to improve institutional profits, and efficiency in this setting also contributes to patient satisfaction. Any misutilization of OR time can prove to be costly. Staffing costs are affected when it leads to overtime pay, canceled cases when a room is behind schedule, and unfilled time translates into missed opportunities for patient care and institutional profit. ■

DISCLOSURES

Name: Rodney A. Gabriel, MD, MAS.
Contribution: This author helped with study design, data curation, statistical analysis, and manuscript preparation.
Conflicts of Interest: The University of California has received funding and product for other research projects from Epimed International (Farmers Branch, TX); Infutronics (Natick, MA); and SPR Therapeutics (Cleveland, OH). The University of California San Diego is a consultant for Avanos (Alpharetta, GA), in which R. Gabriel represents.
Name: Bhavya Harjai, BS.
Contribution: This author helped with study design, statistical analysis, and manuscript preparation.
Conflicts of Interest: None.
Name: Sierra Simpson, PhD.
Contribution: This author helped with study design, statistical analysis, and manuscript preparation.
Conflicts of Interest: None.
Name: Nicole Goldhaber, MD.
Contribution: This author helped with study design and manuscript preparation.
Conflicts of Interest: None.
Name: Brian P. Curran, MD.
Contribution: This author helped with study design and manuscript preparation.
Conflicts of Interest: None.
Name: Ruth S. Waterman, MD, MS.
Contribution: This author helped with study design, data curation, and manuscript preparation.
Conflicts of Interest: None.
This manuscript was handled by: Zeev N. Kain, MD, MBA.

REFERENCES

1. Macario A, Vitez TS, Dunn B, McDonald T. Where are the costs in perioperative care? Analysis of hospital costs and charges for inpatient surgical care. *Anesthesiology*. 1995;83:1138–1144.
2. Casalino LP, Devers KJ, Brewster LR. Focused factories? Physician-owned specialty facilities. *Health Aff (Millwood)*. 2003;22:56–67.
3. Hair B, Hussey P, Wynn B. A comparison of ambulatory perioperative times in hospitals and freestanding centers. *Am J Surg*. 2012;204:23–27.
4. Phieffer L, Hefner JL, Rahmanian A, et al. Improving operating room efficiency: first case on-time start project. *J Healthc Qual*. 2017;39:e70–e78.
5. Cerfolio RJ, Ferrari-Light D, Ren-Fielding C, et al. Improving operating room turnover time in a New York City academic hospital via lean. *Ann Thorac Surg*. 2019;107:1011–1016.
6. Jiao Y, Sharma A, Ben Abdallah A, Maddox TM, Kannampallil T. Probabilistic forecasting of surgical case duration using machine learning: model development and validation. *J Am Med Inform Assoc*. 2020;27:1885–1893.
7. Bartek MA, Saxena RC, Solomon S, et al. Improving operating room efficiency: machine learning approach to predict case-time duration. *J Am Coll Surg*. 2019;229:346–354 e3.
8. Gabriel RA, Waterman RS, Kim J, Ohno-Machado L. A Predictive model for extended postanesthesia care unit length of stay in outpatient surgeries. *Anesth Analg*. 2017;124:1529–1536.
9. Elsharydah A, Walters DR, Somasundaram A, et al. A preoperative predictive model for prolonged post-anaesthesia care unit stay after outpatient surgeries. *J Perioper Pract*. 2020;30:91–96.
10. Pandit JJ, Tavare A. Using mean duration and variation of procedure times to plan a list of surgical operations to fit into the scheduled list time. *Eur J Anaesthesiol*. 2011;28:493–501.
11. Fei H, Meskens N, Chu C. A planning and scheduling problem for an operating theatre using an open scheduling strategy. *Comput Ind Eng*. 2010;58:221–230.
12. Cayirli T, Veral E. Outpatient scheduling in health care: a review of literature. *Prod Oper Manag*. 2003;12:519–549.

13. Bellini V, Guzzon M, Bigliardi B, Mordonini M, Filippelli S, Bignami E. Artificial intelligence: a new tool in operating room management. Role of machine learning models in operating room optimization. *J Med Syst*. 2019;44:20.
14. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22.
15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825–2830.
16. Chawla N, Bowyer K, Hall L., Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–357.
17. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
18. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20:273–297.
19. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26:297–302.
20. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab*. 1948;5:1–34.
21. Dankers F, Traverso A, Wee L, van Kuijk SMJ. Prediction modeling methodology. In: Kubben P, Dumontier M, Dekker A, eds. *Fundamentals of Clinical Data Science*. Springer; 2019:101–120.
22. Nick TG, Campbell KM. Logistic regression. *Methods Mol Biol*. 2007;404:273–301.
23. Connor CW. Artificial intelligence and machine learning in anesthesiology. *Anesthesiology*. 2019;131:1346–1359.
24. Fujiwara K, Huang Y, Hori K, et al. Over- and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis. *Front Public Health*. 2020;8:178.
25. Aceto P, Perilli V, Modesti C, Ciocchetti P, Vitale F, Sollazzi L. Airway management in obese patients. *Surg Obes Relat Dis*. 2013;9:809–815.
26. Reeves JJ, Waterman RS, Spurr KR, Gabriel RA. Efficiency Metrics at an Academic Freestanding Ambulatory Surgery Center: analysis of the impact on scheduled end-times. *Anesth Analg*. 2021;133:1406–1414.
27. van Veen-Berkx E, Elkhuizen SG, van Logten S, et al; Dutch Operating Room Benchmarking Collaborative. Enhancement opportunities in operating room utilization; with a statistical appendix. *J Surg Res*. 2015;194:43–51.e1.
28. Denton B, Viapiano J, Vogl A. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Manag Sci*. 2007;10:13–24.
29. Lebowitz P. Schedule the short procedure first to improve OR efficiency. *AORN J*. 2003;78:651–654.
30. Stepaniak PS, Vrijland WW, de Quelerij M, de Vries G, Heij C. Working with a fixed operating room team on consecutive similar cases and the effect on case duration and turnover time. *Arch Surg*. 2010;145:1165–1170.
31. Dexter F, Dexter EU, Ledolter J. Influence of procedure classification on process variability and parameter uncertainty of surgical case durations. *Anesth Analg*. 2010;110:1155–1163.
32. Pandit JJ, Tavare A. Using mean duration and variation of procedure times to plan a list of surgical operations to fit into the scheduled list time. *Eur J Anaesthesiol*. 2011;28:493–501.
33. Kougias P, Tiwari V, Sharath SE, et al. A statistical model-driven surgical case scheduling system improves multiple measures of operative suite efficiency: findings from a single-center, randomized controlled trial. *Ann Surg*. 2019;270:1000–1004.
34. Eijkemans MJ, van Houdenhoven M, Nguyen T, Boersma E, Steyerberg EW, Kazemier G. Predicting the unpredictable: a new prediction model for operating room times using individual characteristics and the surgeon's estimate. *Anesthesiology*. 2010;112:41–49.
35. Hosseini N, Sir MY, Jankowski CJ, Pasupathy KS. Surgical duration estimation via data mining and predictive modeling: a case study. *AMIA Annu Symp Proc*. 2015;2015:640–648.
36. Stepaniak PS, Heij C, Mannaerts GH, de Quelerij M, de Vries G. Modeling procedure and surgical times for current procedural terminology-anesthesia-surgeon combinations and evaluation in terms of case-duration prediction and operating room efficiency: a multicenter study. *Anesth Analg*. 2009;109:1232–1245.
37. Zhao B, Waterman RS, Urman RD, Gabriel RA. A machine learning approach to predicting case duration for robot-assisted surgery. *J Med Syst*. 2019;43:32.
38. Martinez O, Martinez C, Parra CA, Rugeles S, Suarez DR. Machine learning for surgical time prediction. *Comput Methods Programs Biomed*. 2021;208:106220.
39. Strömblad CT, Baxter-King RG, Meisami A, et al. Effect of a predictive model on planned surgical duration accuracy, patient wait time, and use of presurgical resources: a randomized clinical trial. *JAMA Surg*. 2021;156:315–321.
40. Tuwatananurak JP, Zadeh S, Xu X, et al. Machine learning can improve estimation of surgical case duration: a pilot study. *J Med Syst*. 2019;43:44.