OPEN ACCESS Freely available online

Journal of Anesthesia & Clinical Research

Research Article

Improving and Interpreting Surgical Case Duration Prediction with Machine Learning Methodology

Jesyin Lai[1], Jhao-Yu Huang[1], Shu-Cheng Liu[1], Der-Yang Cho[2], Jiaxin Yu[1*]

[1]AI Innovation Center, China Medical University Hospital, Taichung, Taiwan; [2]Department of Neurosurgery, China Medical University Hospital, Taichung, Taiwan

ABSTRACT

Objective: Hospitals encounter challenges in performing efficient scheduling and good resource management to ensure advanced healthcare quality is provided to patients. Operating room (OR) scheduling is important as it affects workflow efficiency, critical care and OR optimization. Automatic scheduling and accurate surgical case duration prediction have critical roles in improving OR utilization. To estimate surgical case duration, most hospitals rely on historic averages obtained from the electronic medical record (EMR) scheduling systems. However, this produces low accuracy leading to negative impacts, e.g. rescheduling and cancellation.

Methods: A large date set, which covered various details on patients, surgeries, specialties and surgical teams, was obtained. Surgical cases within 60-600 min from 14 specialties were selected for predictive model development. These data included over 500 different procedure types. All models were evaluated with R-square (R2), mean absolute error (MAE), percentage overage (actual duration > prediction), underage (actual duration < prediction) and within. Subsequently, all selected cases were separated into cases with 1 procedure or ≥ 2 procedures and retrained with the best model.

Results: The extreme gradient boosting (XGB) model was superior, achieving a higher R2, lower MAE and higher percentage within on a time-wise testing set (not in the original data). The errors (actual - predictions) could be reduced using model retrained on cases with ≥ 2 procedures (XGB2). Interpretation of XGB predictions with Shapley additive explanations showed that procedure type, anesthesia type, and procedure no. were the top 3 most important features. Specific and higher interactions between anesthesia type, procedure no. and specialty were also identified in a subset of complicated cases.

Conclusions: The XGB and XGB2 models outperformed other models in predicting surgical case durations. They are deployed as a stand-alone machine intelligence server connected by the EMR system for scheduling. This will eventually lead to reduce medical and financial burden for healthcare management.

Keywords: Operating room; Scheduling; Machine learning; Extreme gradient boosting; Anesthesia type; Surgery; Shapley additive explanations

INTRODUCTION

It has become increasingly important for clinics and hospitals to manage resources for critical care during the COVID-19 pandemic period. Statistics show that approximately 60% of patients admitted to the hospital will need to be treated in the operating room (OR) [1], and the average OR cost is up to 2,190 dollars per hour in the United States [2,3]. Hence, the OR is considered as one of the highest hospital revenue generators and accounts for as much as 42% of a hospital's revenue [3,4]. Based on these statistics, a modern OR scheduling and management strategy is not only critical to patients who are in need of elective, urgent and emergent surgeries but is also important for surgical teams to be prepared. Owing to the high importance of the OR, OR efficiency improvement has high priority so that the cost and time spent on the OR is minimized while the OR utilization is maximized to increase the surgical case number and patient access [5].

In a healthcare system, numerous factors are involved in affecting OR efficiency, for example, patient expectation and satisfaction, interactions between different professional specialties, unpredictability during surgeries, surgical case scheduling, etc. [6]. Although the OR process is complex and involves multiple parties, one way to enhance OR efficiency is by increasing the accuracy of

Correspondence to: Dr. Jiaxin Yu, AI Innovation Center, China Medical University Hospital, Taichung, No. 2, Yude Road, North District, Taichung, Taiwan; E-mail: jiaxin.yu@mail.cmuh.org.tw

Received: April 01, 2021; **Accepted:** April 15, 2021; **Published:** April 22, 2021

Citation: Lai J, Huang JY, Liu SC, Cho DY, Yu J (2021) Improving and Interpreting Surgical Case Duration Prediction with Machine Learning Methodology. J Anesth Clin Res. 12:998.

predicted surgical case duration. Over- or under-utilization of OR time often leads to undesirable consequences such as idle time, overtime, cancellation or rescheduling of surgeries, which may induce a negative impact on the patient, staff and hospital [7]. In contrast, high efficiency in OR scheduling not only contributes to a better arrangement for the usage of the OR and resources but can also lead to a cost reduction and revenue increase since more surgeries can be performed.

Currently, most hospitals schedule surgical case duration by employing estimations from the surgeon and/or averages of historical case durations, and studies show that both of these methods have limited accuracy [8-10]. For case lengths estimated by surgeons, factors including patient conditions and anesthetic issues might not be taken into consideration. Moreover, underestimation of case duration often occurs because surgeon estimations are usually made by favoring maximizing block scheduling to account for potential cancellations and cost reduction. Furthermore, operations with higher uncertainty and unexpected findings during surgery add difficulties and challenges to case length estimation [8]. Historical averages of case duration for a specific surgeon or a specific type of surgery obtained from electronic medical record (EMR) scheduling systems have also been used in hospitals. However, these methods have been shown to produce low accuracy due to the large variability and lack of the same combination of factors in the preoperative data available on the case that is being performed [11].

To improve the predictability, researchers have utilized linear statistical models, such as regression, or simulation for surgical duration prediction and evaluation of the importance of input variables [10,12–14]. However, a common shortcoming of these studies is that relatively fewer input variables or features were used in their models than in alternative approaches due to the limitation of statistical techniques in handling too many input variables. Recently, machine learning (ML) has been shown to be powerful and effective in aiding health care management. Master et al. (2017) trained multiple ML models, including decision tree regression, random forest regression, gradient boosted regression trees and hybrid combinations, to automate prediction and classification of pediatric surgical durations [15].

Ensemble algorithms, implementing least-squares boosting and bagging models with ML, developed by Shahabikargar *et al.* were shown to reduce the error by 55% compared to the original error [7]. With the use of a boosted regression tree, Zhao *et al.* increased the percentage of accurately booked cases for robot-assisted surgery from 35% to 52%. Bartek *et al.* reported that they were able to improve predicted cases within 10% of the tolerance threshold from 32% to 39% using an extreme gradient boosting model [16]. Nonetheless, these ML studies included only 5-12 different types of procedures and specialties to train their ML models, which may limit the generalization of these models.

In this study, more than 170,000 cases were obtained from China Medical University Hospital (CMUH) containing hundreds types of procedures across multiple different specialties. From the original data, we analyzed the working time of primary surgeons and computed their total number of previous surgeries and the total time spent on previous surgeries within 24 hr as well as within the last 7 days. Since surgeons' working performance might be affected by previous events, surgical cases performed by the same primary surgeon should not be considered as totally independent and unrelated. Hence, previous surgical counts and working time obtained from surgeons' data were included as additional features in our ML model training to account for their influences on surgical case duration. With a total of 20 features, ML models were built and trained to improve surgical case duration prediction. Subsequently, model predictions were interpreted with Shapley additive explanations (SHAP) to unravel global importance of features in the ML model and local interaction of features in a subset of cases.

METHODS

Data sources

Data for this study were collected retrospectively from the EMR scheduling system of CMUH located in Taichung, Taiwan. The data set covered a broad variety of details about patients, surgeries, specialties and surgical teams. A total of 170,748 cases performed between Jan 1, 2017, and Dec 31, 2019, were used for model development. Additionally, 8,672 cases performed between Mar 1 and April 30, 2020, were used as data for time-wise model evaluation in this study. The proportions regarding patient characteristics in the overall data set and time-wise testing set were reported in Table 1. Over 500 different types of procedures across 14 surgical specialties were included in the training data set. Institutional review board approval (CMUH109-REC1-091) was obtained from CMUH before carrying out this study.

Selection and exclusion criteria, data processing and feature selection

Emergent and urgent surgical cases were removed since these two types of surgeries cannot be scheduled until they happen. Surgical records with missing values in procedure type and specialty were excluded. Patients who were pregnant, patients' age younger than 20 and duplicates were also removed. The exclusion criteria are shown in Figure 1. Since surgical cases that

Table 1: Proportion based on patient characteristics in the overall original data set and time-wise testing set.

	Overall (n = 86,621)	Test(n =4,257)
Gender		
Male	43935 (50.7%)	2105 (49.4%)
Female	42686 (49.3%)	2152 (50.6%)
Age		
20-45	24740 (28.6%)	1188 (27.9%)
45-65	33829 (39.1%)	1699 (39.9%)
65-80	21481 (24.8%)	1054 (24.8%)
>80	5663 (6.5%)	280 (6.6%)
In-/out-patient		
In-patient	73585 (85%)	3681 (86.5%)
Out-patient	13036 (15%)	576 (13.5%)
Hypertension		
Yes	26039 (30%)	1287 (30.2%)
No	49059 (56.6%)	2427 (57%)
Unknown	11523 (13.3%)	543 (12.8%)
BMI		
<18.5	21164 (24.4%)	1176 (27.6%)
18.5-22.9	28466 (32.9%)	1486 (34.9%)
23-26.9	26122 (30.2%)	1313 (30.8%)
>= 27	5065 (5.8%)	253 (5.9%)
Missing	5804 (6.7%)	29 (0.7%)

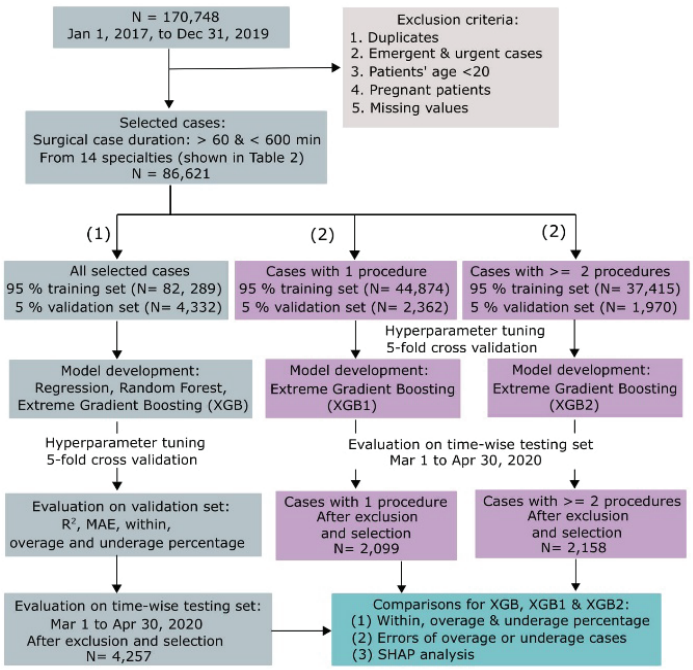

Figure 1: The workflow of model training and evaluation. The data used for model training fall within the time range of Jan 1, 2017, to Dec 31, 2019. From this data set, cases were excluded based on the following criteria: duplicates, emergent and urgent cases, patients with age younger than 20, pregnant patients, and cases with missing values. The total number of cases included in the data set for model development was 86,621. There were two stages (labeled as (1) and (2)) of machine learning (ML) model development in this study. In stage 1, the selected data were split into training (95 %) and validation (5 %) subsets. ML and linear regression models were developed on the training data set and evaluated on the validation data set using R 2 and mean absolute error (MAE). Model performance was also validated using percentage within (absolute duration differences falling within 0.15 x prediction or 30 min), overage (actual > prediction) and underage (actual < prediction).

consumed longer duration may benefit more from predictions using ML approach and human interpretation may fail to consider all the complex interaction of the variables, surgical cases within duration of 60-600 min were selected. Surgical cases that took more than 10 hours were not considered as the case numbers were low. These exclusion and selection criteria resulted in a data set of 86,621 cases that were used for model training and validation. The same criteria were also applied to the data of Mar 1 to April 30, 2020, and 4,257 cases remained after exclusion and selection. These data were treated as time-wise testing set since they were temporally segregated from the original data.

Features were selected from available data sources based on literature review and discussion with surgeons and administrators of CMUH. Although the model performance could be enhanced by some postoperative information (e.g., total blood loss), these parameters cannot be used as features for model training because they were either missing or simply estimated by surgeons before surgery. Therefore, only variables that were available before surgery were selected for model development. Furthermore, the correlations of feature variables with the surgical case duration were checked by performing a regression analysis. Only those variables with significant (p-value < 0.05) correlation coefficients were selected as predictor variables for model training. When visualizing all the categories of procedure type and the International Classification of Diseases (ICD) code, there were hundreds to thousands of categories in these two variables. To reduce the problem of having too many dimensions during one-hot encoding of categorical features and running into the curse of dimensionality, procedure type and ICD codes with less than 30 cases within 2017-2019 were not included in one-hot encoding. Surgical cases with these rare procedure types were removed as well. In addition, since surgical case duration can be related to the performance of surgeons and surgeons' performance is affected by their working time, we analyzed primary surgeons' previous surgical events. The number of previous surgeries and total surgical minutes performed by the same primary surgeons within the last 7 days and 24 hr. Together, 20 predictor variables were included for predictive model development in this study. These predictors can be categorized into 5 groups: patient, surgical team, operation, facility and primary surgeon's prior events (Table 2).

Model development and training

We trained multiple algorithms for surgical case duration prediction. Surgical case duration (in minutes) is the total period starting from the time the patient enters the OR to the time of exiting the OR ("wheels-in to wheels-out"). The distribution of surgical case duration was observed to be skewed to the right and it follows a log-normal distribution as reported by some past studies [9,16]. A logarithmic transformation on the surgical case duration was performed. All models were built by using the log-transformed

Table 2: Preoperative data with 20 predictor variables were used as inputs for model development. The predictor variables can be categorized by relationship to patient, surgical team, operation, facility and surgeon's prior events. These predictor variables were selected based on the significance (p-value<0.05) of their correlations with the outcome using a regression analysis. Text in parentheses is the code name of the corresponding feature variable that was used during model development and interpretation. BMI classification was performed following the standard for Asians and the categories are listed in Table 1.

Patient	Surgical team	Primary surgeon's prior events
Age	Primary surgeon's ID (DocID)	No. of previous surgeries performed by the surgeon on the same day (Opcount_1d)
Gender (SexName)	Surgeon team size (TeamSize)	Total surgical hours performed by the surgeon on the same day (Optoltime_1d)
ICD code (Diag)	Specialty (DivNo)	No. of previous surgeries performed by the surgeon within the last 7 days (Opcount_7d)
In-/out-patient (OpType)	Primary surgeon's age (Dr_age)	Total surgical hours performed by the surgeon within the last 7 days (Optoltime_7d)
ASA status (ASA) Hypertension	Primary surgeon's year of eperience (Dr_year)	
Operation	**Facility**	
Procedure type (Proceed)	Room No. (OpRoom)	
Anesthesia type (Ana Value)	Day of the week (weekday) Time of day (Time of Day)	

ICD: International Classification of Diseases; ASA: American Society of Anesthesiologists; BMI: Body mass index

case duration as the target. Data visualization, processing, model development and evaluation in this study were all performed using Python.

There were two stages (stage 1 and stage 2 shown in Figure 1) of model development in this study.

Moreover, the models were further evaluated on the most recent surgical cases (from Mar 1 to Apr 30, 2020), which were not included in the original data set for model training. In stage 2, all selected cases were divided into cases with 1 procedure and ≥ 2 procedures. These two groups of cases were then retrained with XGB algorithm since XGB produced the best predictive performance in stage 1. The same strategies, including the train-valid split ratio and hyperparameter tuning with 5-fold cross validation, for model development in stage 1 were applied in stage 2. Two additional ML models were built separately to predict surgical case durations for cases with 1 procedure (XGB1) or ≥ 2 procedures (XGB2). Eventually, comparisons were performed on the XGB model (stage 1), XGB1 and XGB2 (stage 2) in the following three aspects: (1) within, overage and underage percentage; (2) errors of cases that were categorized as overage or underage; (3) feature importance of model revealed in SHAP analysis. SHAP: Shapley additive explanations.

A data-splitting strategy was used in the training for all the models. In stage 1, the original data were randomly split into training and validation subsets at a ratio of 95%: 5%. The training data were used to build different predictive models as well as to extract important predictor variables. The validation data were used for internal evaluation of the models. In addition to interval evaluation, time-wise evaluation on all the models was performed using data from Mar 1 to Apr 30, 2020. These data were not included in the original data set for model training. The results obtained from time-wise evaluation are better in verifying the robustness of the trained model in making an accurate prediction since they were temporally segregated from the original data. Moreover, by using train-valid split ratio of 95%: 5%, the data size of the validation set was similar to the data size of the testing set. Historical averages of case durations based on procedure-specific data obtained from EMR systems were used as the baseline model for comparison. A multivariate linear regression (Reg) model was built to be used as a model for comparison. Two ML algorithms were trained to predict surgical durations in this study. The first ML algorithm used is random forest (RF), a tree-based supervised learning algorithm. RF uses bootstrap aggregation or a bagging technique for regression by constructing a multitude of decision trees based on training data and outputting the mean predicted value from the individual trees [17]. Tree-based techniques were suitable for our data since they include a large number of categorical variables, e.g., ICD code and procedure type, of which most were sparse. Extreme gradient boosting (XGB) algorithm is the second ML algorithm that was trained for comparison to the RF model. Recently, XGB algorithm has gained popularity within the data science community d ue to its ability in overcoming the curse of dimensionality as well as capturing the interaction of variables [18]. XGB is also a decision tree-based algorithm similar to RF. XGB and RF algorithms are different in the way on how the trees are built. It has been shown that XGB performs better than RF if parameters are tuned carefully [19,20]. For both RF and XGB algorithms, we adopted a 5-fold cross-validation strategy to tune the best hyperparameters, e.g., no. of estimators, maximum of depths, etc.

In stage 2, all the selected cases from the original data as well as the time-wise testing set were divided into cases with 1 procedure and ≥ 2 procedures. This is because we observed that approximately 55% of the selected cases in the original data (47,236 cases) were cases with 1 procedure while the remaining consisted of cases with ≥ 2 procedures. Moreover, cases with 1 procedure were mostly less complicated than cases with ≥ 2 procedures. Building two separate ML models that predict durations for cases with 1 procedure or ≥ 2 procedures might help to improve overall predictive performance. Therefore, these two groups of cases were retrained with the ML algorithm that produced the best results in stage 1. The same strategies, including train-valid split ratio and hyperparameter tuning with 5-fold cross validation, for model development in stage 1 were applied in stage 2. Cases with 1 or ≥ 2 procedures in the training set were used to train model to predict surgical case durations for cases with 1 (XGB1) or ≥ 2 procedures (XGB2).

At last, three types of comparisons were performed on the XGB (stage 1), the XGB1 and XGB2 (stage 2) models: Within, overage and underage percentage; Errors of cases that were categorised as overage or underage; Feature importance of model revealed in Shapley additive explanations (SHAP) analysis.

Model evaluation

The three key metrics used to evaluate model performance in this study included (1) R-square (R^2), (2) meanb absolute error (MAE), and (3) the percentage within, overage, and underage. R^2 is the coefficient of determination representing the proportion of the variance for the actual case duration that is explained by predictor variables in the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

In the above and below equations, y is the actual case duration, $\hat{y}$ is the predicted case duration and $\overline{y}$ is the mean of all actual durations. Meanwhile, MAE measures the average of errors between the actual case durations and the predictions.

$$MAE = \frac{\sum_{i=1}^{N} | y_i - \hat{y}_i |}{N}$$

In this study, percentage within indicates the percentage of cases with absolute duration differences falling within a tolerance threshold ($\tau(y)$). Two types of tolerance threshold were used: (1) $\tau(y)$= 0.15 x prediction and (2) $\tau(y)$= 30 min. Meanwhile, percentage underage is the percentage of cases with actual case duration shorter than prediction and case duration difference was more negative than the threshold. Similarly, percentage overage is the percentage of cases with actual case duration longer than prediction and case duration difference was more positive than the threshold. The condition that defines a case as overage, within and underage is summarised as follows:

$$\text{condition} = \begin{pmatrix} \text{overage}, & y_i - \hat{y}_i & \geq \tau(\hat{y}) \\ \text{within}, & | y_i - \hat{y}_i | & < \tau(\hat{y}) \\ \text{underage}, & y - \hat{y} & \leq -\tau(\hat{y}) \end{pmatrix}$$

Model interpretation with SHAP

To further interpreter the developed XGB model, we determined the SHAP value by applying SHAP package. SHAP can be used

to explain the predicted output by computing the contribution of each feature to the prediction [21]. SHAP value of a feature i, termed ϕi, can be obtained with the following equation:

$$\phi i = \sum_{S \subseteq F\{i\}} [\frac{|S|!(|F| - |S| - 1)!}{|F|!}][f(S \bigcup \{i\} - f(S))$$

In the above equation, F is the set of all features considered for the XGB algorithm, S denotes a subset of features obtained from the set F except feature i, and f(S) is the expected output given by the set S of features. In summary, SHAP values indicate the impact of a feature on the model output. For our ML model, a large positive (negative) SHAP value of a feature implies that this feature has a large contribution in predicting a longer (shorter) surgical case duration. Meanwhile, a SHAP value of 0 implies that this feature have no or low contribution in predicting surgical case duration. SHAP values are expressed in log-odds (5) in this study

$$\log[P(\phi)/(1 - P(\phi))],\ P(\phi) < 1 \cdot$$

Local interaction effects between features were identified by applying SHAP interaction values in the SHAP package. While SHAP value is the attribution for each feature, SHAP interaction value is a matrix of feature attributions [22]. The interaction effects on the off-diagonal and the main effects are on the diagonal. The SHAP interaction values is defined as:

$$\sum_{S \subseteq m/i,j} [\frac{||S|!(M - |S| - 2)!}{2(M-1)!} \Delta ij(f,\ x,\ S)$$

when i ≠ j and:

$$\Delta ij(f,\ x,\ S) = fx(S \bigcup \{i,\ j\}) - fx(S \bigcup \{i\}) - f_x(S \bigcup \{j\}) + f_x(S)$$

In the equation (6), m is the set of all M input features. More details regarding the computation 180 of SHAP interaction values can be referred to Lundberg et al. (2020). The SHAP interaction value between feature i and feature j is split equally between each feature ($\Phi_{i,j}(f, x) = \Phi_{j,i}(f, x)$), and the total interaction effect is the sum of $\Phi_{i,j}(f, x)$ and $\Phi_{j,i}(f, x)$.

RESULTS

Model development and evaluation

Since surgical cases completed within one hour can be estimated by surgeons easily, predictions made by ML model would provide more benefits to surgical cases longer than 60 min. On the other hand, surgical cases conducted more than 10 hr (60 min) were rare. Therefore, surgical cases with actual duration longer than 60 min and less than 600 min were selected from the original data from Jan 1, 2017, to Dec 31, 2019. Fourteen specialties with surgical cases numbers ≥ 100 per month were included in this study. These specialties are shown in Table 2. After processing data with the exclusion and selection criteria as shown in Figure 1, 82,289 cases containing 546 procedural categories were included for predictive model development. Furthermore, a recent data set collected from Mar 1 to April 30, 2020, was processed similarly to the original data set and used as the time-wise testing set to verify the robustness of model performance. There were total 4,257 cases in the time-wise testing set. In the model development of stage 1, the original processed data were split into training and validation sets at 95%:5%. This generated a validation set with a case number of 4,332, which is close to the case number of the time-wise testing set.

Stage 1

In the hospital, surgical cases are scheduled according to estimates made by primary surgeons. However, surgeon estimates rely heavily on prior experiences, which are based on the procedures that have been performed. Since there is no formal record on surgeon estimates, an average model built with the procedure type only was used as the baseline model, termed average procedure-specific model, in stage 1. This average procedure-specific model closely reflects the scenario of surgeon estimates used by the hospital. All built models were evaluated with R-square (R2), mean absolute error (MAE), percentage overage (actual > prediction), underage (actual < prediction) and within (absolute duration differences falling within a threshold) on the validation set and time-wise testing set, respectively. The results of R2 and MAE are reported in Table 4. The average procedure-specific model had a R2 value of 0.68 and a MAE of 41.3 min on the time-wise testing set. Since no other feature was taken into consideration in this baseline model except the procedural duration of surgical cases that happened in the past, it exhibited higher prediction error and lower accuracy. A linear regression (Reg) model was built by including 20 input variables shown in Table 3.

The R2 value increased to 0.72 and the MAE decreased to 38.2 min on the time-wise testing set. This indicates that predictive performance of the model improved when other information was taken into consideration. However, Reg model is still not complicated enough to consider various types of interactions between input variables in a real-world situation. ML algorithms are helpful in making predictions in a more complicated scenario. The random forest (RF) model and extreme gradient boosting (XGB) model were trained subsequently to improve predictions.

Table 3: Surgical cases from the above specialties were used to develop ML models to predict surgical case durations.

Specialties included for model development
1. Body science and metabolic disorders
2. Trauma and acute care surgery
3. General surgery
4. Orthopedics
5. Urology
6. Neurosurgery
7. Colorectal surgery
8. Thoracic surgery
9. Obstetrics and gynecology
10. Otorlaryngology, head and necy surgery
11. Ophthalmology
12. Plastic and reconstruction surgery
13. Cadiovascular surgery
14. Breast surgical oncology

Table 4: The results of R-square (R2) scores and mean absolute errors (MAE) of multiple models evaluated on validation and time-wise testing sets. Procedure: average procedure-specific model; Reg: regression; RF: random forest; XGB: extreme gradient boosting.

	Validation set				Time-wise testing set			
Metric	Procedure	Reg	RF	XGB	Procedure	Reg	RF	XGB
R2	0.68	0.72	0.74	0.77	0.68	0.72	0.74	0.77
MAE (min)	38.9	36.4	32.8	31.7	41.3	38.2	36.5	34.5

The performance of both the RF and XGB models was better than the Reg and the baseline models. For the RF model, the R2 value was 0.74 and the MAE was 36.5 min. For the XGB model, the R2 value was 0.77 and the MAE was 34.5 min. Among all the trained models, the XGB outperformed the others in terms of R2 and MAE values. Moreover, the performance of the XGB on the time-wise testing set was similar to the validation set (R2=0.77 and MAE=32.8 min) even though the time-wise testing set (March-April, 2020) was temporally segregated from the training and validation sets (2016-2019). This also reflects that the XGB model generalized well during the pandemic period. For percentage within, two criteria were used to define the tolerance threshold: 0.15 x prediction or 30 minutes. In both criteria, percentages of overage and underage were decreased while percentage within was increased when comparing the XGB model to the baseline model (Figure 2).

When 0.15 x prediction was used as the threshold, there was an 18% increase in percentage within for the XGB model compared to the baseline model. Meanwhile, a 10% increase in percentage within was observed in the XGB model compared to the baseline model when 30 min was used as the threshold.

Bland-Altman (BA) plots using the time-wise testing set for the baseline model and the XGB model are shown in Figure 3. The BA plots clearly show that predictions generated by the XGB model had smaller and less scattered errors (actual - prediction) compared to the average model of procedure for the XGB model, the range of ±1.96 x standard deviation for errors (red dashed lines) is also narrower.

These demonstrates that the XGB model is more accurate than the baseline model in predicting surgical case duration. Overall, the XGB model had the best performance. Hence, subsequent analysis an interpretation were focused on the XGB model. The XGB algorithm was then used to build models that predict durations for cases with 1 or ≥ 2 procedures in stage 2.

Stage 2

Model development in stage 2 was conducted in attempt to test if building two separate predictive models to predict durations for

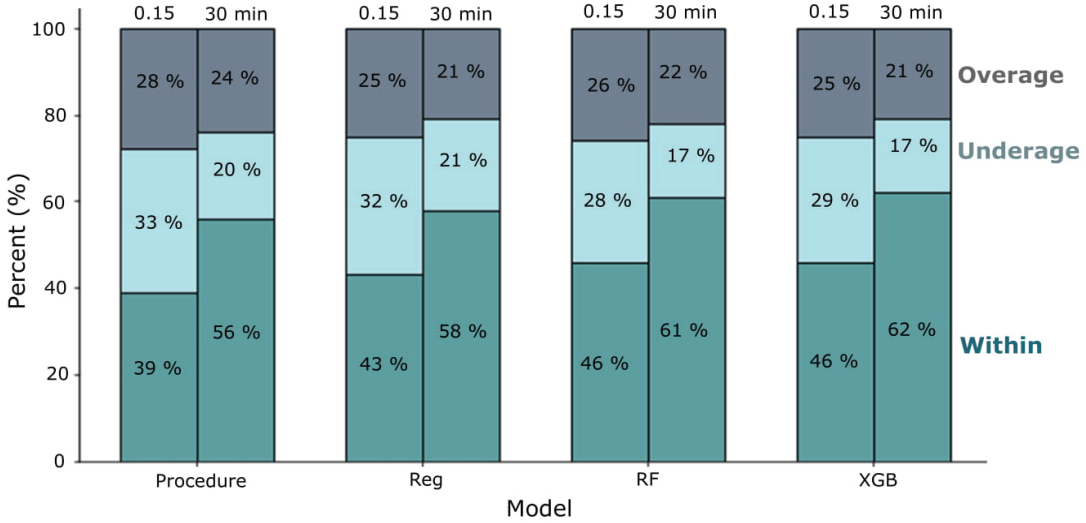

Figure 2: Machine learning algorithms improved predictions of surgical case duration. The performance of all models was evaluated on a time-wise testing set (data not included in the original data set for model training) by using percentage overage (actual duration longer than prediction), underage (actual duration shorter than prediction) and within (absolute duration differences falling within a threshold). Two criteria were used to define the tolerance thresholds for predictions to be considered as within: 0.15 x prediction or 30 minutes. For both criteria, percentages of overage and underage were decreased while percentage within was increased when comparing the extreme gradient boosting (XGB) model to the average model for procedure. Procedure: average procedure-specific model; Reg: regression; RF: random forest; XGB: extreme gradient boosting.

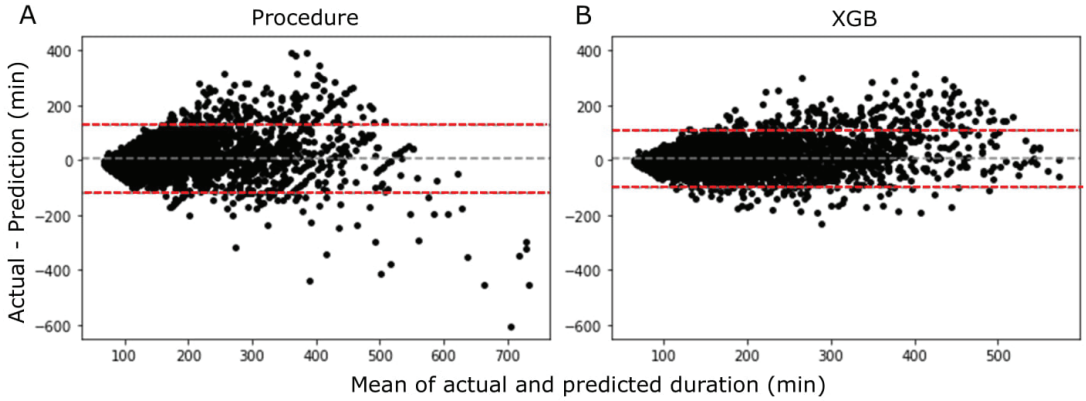

Figure 3: Predictions produced by the extreme gradient boosting (XGB) model were closer to actual surgical case durations compared to predictions produced by the average procedure-specific model. The Bland-Altman plots of (A) the average procedure-specific model and (B) the XGB model reveals that predictions made by the XGB model had smaller and less scattered errors (actual - prediction) based on the results of the time-wise testing set. The range of ±1.96 x standard deviation for errors (indicated by red dashed lines) is also narrower for predictions generated by the XGB model compared to the average model.

cases with 1 procedure or ≥ 2 procedures would further improve predictions. When focusing on the time-wise testing set, the two ML models, XGB1 and XGB2, produced similar within, overage and underage percentages as the XGB model regardless of whether the tolerance threshold of 0.15 X prediction or 30 min was used. Although the percentages were similar, the prediction errors (actual - prediction) made by the XGB, the XGB1 and XGB2 models might be different. Hence, we compared the prediction errors for cases which were categorized as overage or underage, respectively. Prediction errors of the XGB1 model were compared to prediction errors of the XGB model on the cases with 1 procedure extracted from the time-wise testing set. Similarly, prediction errors of the XGB2 model were compared to prediction errors of the XGB model on the cases with ≥ 2 procedures extracted from the time-wise testing set. Prediction errors for overage or underage as well as the respective percentages of outliers were analyzed and compared [4]. For the XGB1 model, both prediction errors and the percentages of outliers were not improved compared to XGB model (results not shown). For the XGB2 model, the percentages of outliers for overage were similar to the XGB model but the percentages of outliers for underage were reduced compared to the XGB model (Figure 4). In addition, the prediction errors of the outliers for underage were significantly lower (p-value > 0.05, non-parametric ranksum test) in the XGB2 model compared to those of the XGB model when using the tolerance threshold of 30 min. As XGB2 were able to reduce prediction errors for cases with ≥ 2 procedures, the XGB model will be used to predict surgical case durations for cases with 1 procedure while the others will be predicted by the XGB2 model.

Model interpretation

To uncover the global importance and the impact of each feature on the XGB, XGB1 and XGB2 model output, we applied SHAP to explain the model. SHAP adopts the classical Shapley values estimation methods, which satisfy the desirable properties of local accuracy, missingness and consistency [21,23]. It explains model output by computing the contribution of each feature to the prediction. Figure 5A shows feature impact on model output based on SHAP value while Figure 5B-D shows the top 20 one-hot encoded features with the largest average SHAP value magnitude for the XGB, XGB1 and XGB2 model, respectively. Procedure type, anesthesia type, no. of procedure, specialty and in-/out-patient were the 5 most important features in the XGB model (Figure 5).

Specific code names are used to represent the full names of features in the figure. The features' full names can be referred to Table 3. Features labeled with red boxes were the features (belong to anesthesia type or procedure type) contributed to longer duration while features labeled with black boxes contributed to shorter duration. (C) The top 20 one-hot encoded features with the largest average SHAP value magnitude in the XGB1 model. (D) The top 20 one-hot encoded features with the largest average 31 SHAP value magnitude in the XGB2 model. AnaValue_GA: general anesthesia; Ana Value_LA: local anesthesia; AnaValue_IG: intravenous anesthesia; P_83046: spinal fusion with spinal

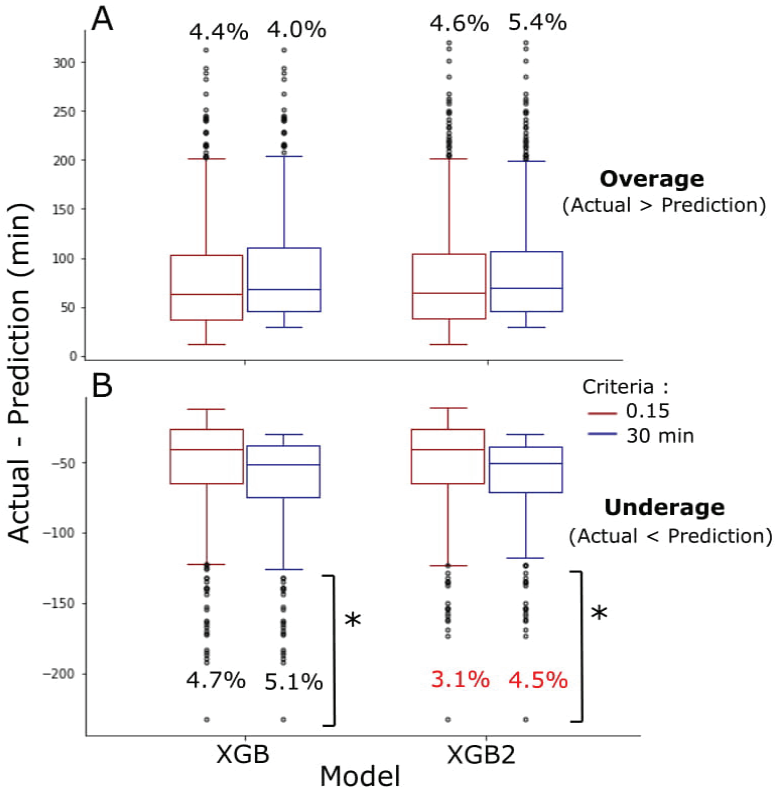

Figure 4: The percentages (red text) and the range (*) of errors (actual - prediction) for outliers were reduced in predictions made by the XGB2 model compared to the XGB model for underage cases with ≥ 2 procedures. Errors between actual and predicted durations were compared among the XGB, XGB1 and XGB2 models. Predictions made by the XGB1 model were not significantly improved compared to the XGB model and were not included in this figure. Errors of cases categorized as overage (A) or underage (B) for ≥ 2 procedures were plotted using boxplots. Cases were categorized as overage/underage when their actual durations were either 15 % (red) or more than 30 min (blue) longer/shorter than predicted durations. The percentages shown in the boxplots indicate the percentages of outliers in different conditions. In the XGB2 model, the percentages of underage outliers were reduced. For underage cases and actual durations more than 30 min shorter than predicted durations, the errors of outliers were also significantly reduced (* p-value < 0.05, nonparametric ranksum test) compared to those of the XGB model.

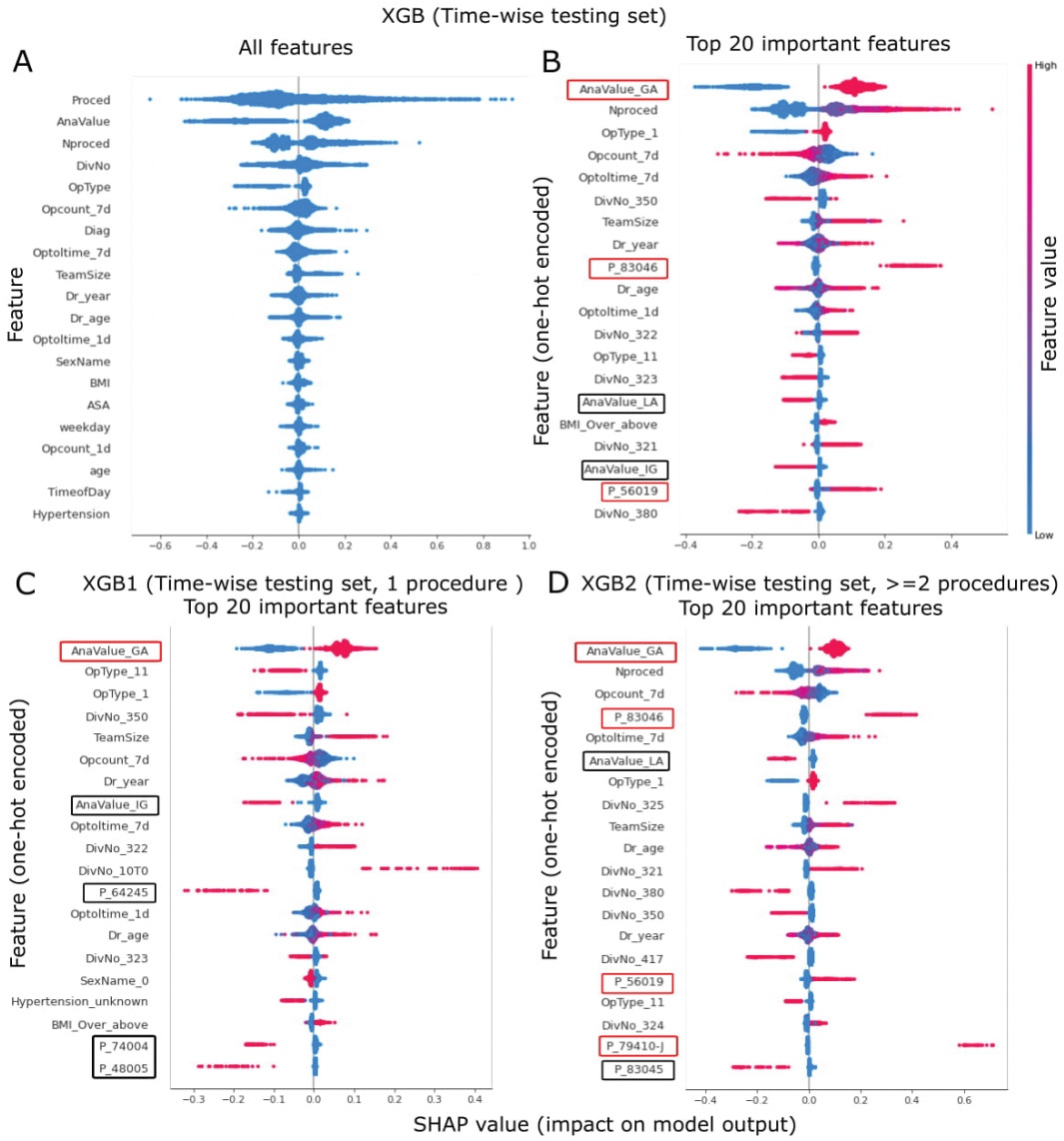

Figure 5: Procedure type, anesthesia type, no. of procedure, specialty and in-/outpatient were the top 5 important features used by the extreme gradient boosting (XGB) model to make predictions. Features were arranged and ranked in descending order according to SHAP value magnitude. (A) SHAP values of all categories within a categorical feature (e.g., procedure type, anesthesia type, specialty, in-/out-patient, ICD code, gender, BMI category, ASA, day of the week, time of day and hypertension) were added to obtain the total SHAP value for that categorical feature. This reveals the total importance for each of the 20 features in the XGB model. (B) The top 20 onehot encoded features with the largest average SHAP value magnitude in the XGB model. Each dot corresponds to a case in the time-wise testing set. Features with long positive tails tended to increase model output (longer surgical case duration) when feature values were either high (for numerical feature) or 1 (for categorical feature).

instrumentation ≤ 6 segments; P_56019: brain and spinal surgery applying microscope; P_64245 removal of internal fixator; P_74004: laparoscopic appendectomy; P_48005: Debridement (5-10 cm); P_79410-J: radical prostatectomy with bilateral pelvic lymph node dissection; P_83045: spinal fusion without spinal instrumentation.

Notably, 3 of the top 5 important variables were attributed to operative information (i.e., procedure type, anesthesia type and no. of procedure). Moreover, two of the features that we computed from the surgeon data (i.e., total surgical minutes and the no. of previous surgeries performed by the surgeon within the last 7 days) had important contributions to model output as they were included within the top 10 list of feature importance. SHAP value was applied in this study to reveal feature importance because it is informative in providing quantification and visualization of the impact (negative or positive) of each feature on the final output for each case as well as the variations in feature contribution relative to changes in feature value.

Since procedure type and anesthesia type were the two most important feature in the XGB model, we focused on the importance and impact of the categories, i.e. one-hot encoded features, under these two critical features. In Figure 5B-D, one-hot encoded features related to anesthesia type or procedure type that contributed to longer case durations (positive impact) are labeled with red boxes while those that contributed to shorter case durations (negative impact) are labeled with black boxes. In all the three models (XGB, XGB1 and XGB2), general anesthesia (GA) was the most important feature that contributed to longer case durations. In contrast, local anesthesia (LA) and intravenous anesthesia (IG) contributed to shorter durations. IG had higher negative impact in cases with 1 procedure (XGB1) while LA had higher negative impact in cases with ≥ 2 procedures (XGB2). For procedure type, those that contributed to shorter durations were mostly revealed in the top 20 important features of XGB1 (5C). Removal of internal fixator (code: 64245), laparoscopic appendectomy (74004) and

debridement of 5-10 cm (48005) had negative impact on surgical case durations in cases with 1 procedure. Meanwhile, spinal fusion without spinal instrumentation (83045) had negative impact on the case durations of cases with ≥ 2 procedures. Contrarily, spinal fusion with spinal instrumentation ≤ than 6 segments (83046), brain and spinal surgery applying microscope (56019), and radical prostatectomy with bilateral pelvic lymph node dissection (79410-J) had positive impact on the model outcome of XGB2 (5D).

While SHAP summary plot provides global interpretability reflecting the general behavior of the features in the model, the specific behavior of the features in a subset of model predictions was explored on complex cases. Surgical cases with procedure numbers ≥ 2, team size ≥ 2 and in-patient cases were considered as complex cases. This is because complex cases usually contain multiple procedure types in a case and involve more surgeons. Meanwhile, in-patient cases usually involve patients with more complications and serious conditions. These complex cases were extracted from the time-wise testing set for local interpretability in the XGB model (Figure 6).

Surgeons with shorter total surgical time and performed more surgical cases in a week were associated with shorter surgical case durations. (E) SHAP interaction value for no. of procedure and procedure type of brain or spinal surgery that used microscope (P_56019). Brain or spinal surgeries that used microscope were associated with shorter durations for cases with ≥ 3 procedure numbers. (F) SHAP interaction value for no. of procedure and procedure type of spinal fusion with spinal instrumentation ≤ than 6 segments (P_83046). Regardless of the no. of procedure, longer durations were consumed for cases with this procedure code.

A heatmap showing SHAP interaction values of one-hot encoded features for the extracted complex cases. One-hot encoded features with top 12 largest SHAP interaction values are reported in the figure. Features with higher and interesting interactive effects were highlighted in red boxes and their SHAP interaction values were plotted in Figure 6B-F. From these SHAP interactive value plots, a few interesting phenomena were observed. In general, complex surgical cases that used GA consumed longer duration than cases that used other anesthesia types. However, surgical cases using GA from the Orthopedics specialty were shorter than GA cases from other specialties. Total surgical minutes ('Optoltime_7d') interacted strongly with total no. ('Optolcount_7d') of previous surgeries performed by the surgeon within the last 7 days. Surgeons with higher

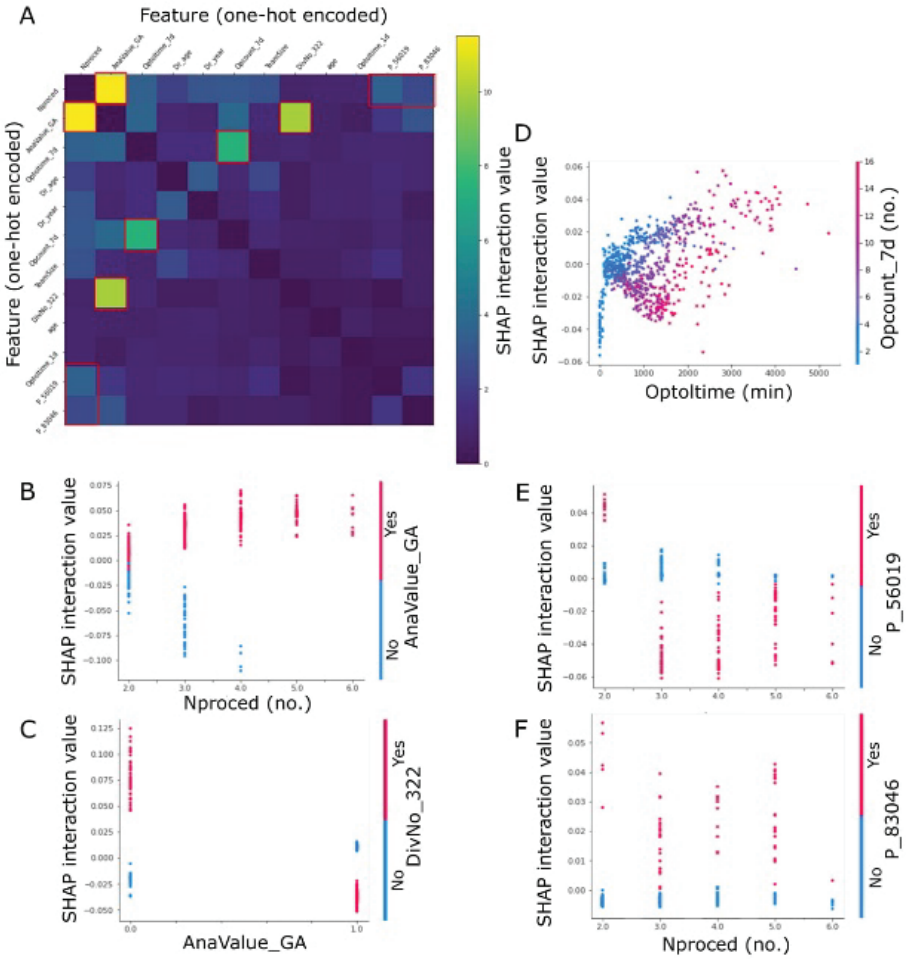

Figure 6: Shapley additive explanations (SHAP) interaction value disclosed interactions between various one-hot encoded features for more complex cases (i.e., procedure numbers ≥ 2, team size ≥ 2 and in-patient cases). (A) A heatmap summarizing SHAP interaction values of one-hot encoded features for complex cases extracted from the testing set. One-hot encoded features with top 12 largest SHAP interaction values are shown. Note that the interaction effects are shown off-diagonal and symmetrically. Higher and interesting interaction effects are highlighted with red boxes. (B) SHAP interaction value for no. of procedure and anesthesia type of general anesthesia (AnaValue_GA). Surgical cases that used GA consumed longer duration and most surgical cases with 5-6 procedure numbers were performed with patients under GA. (C) SHAP interaction value for anesthesia type of GA and Orthopedics (DivNo_322). In general, surgical cases using GA were longer than those using other anesthesia types, but surgical cases using GA from Orthopedics specialty were shorter than others. (D) SHAP interaction value for total surgical minutes (Optoltime_7d) and no. (Opcount_7d) of surgeries performed by the surgeon within the last 7 days.

'Optoltime_7d' and higher 'Optolcount_7d' were associated with longer case durations. However, surgeons with moderate 'Optoltime_7d' and higher 'Optolcount_7d' were associated with shorter case durations. For brain or spinal surgeries that used microscope, they were associated with shorter durations for cases with ≥ 3 procedure numbers. For procedure type of spinal fusion with spinal instrumentation ≤ than 6 segments, the presence of this procedure increased surgical case durations for different no. of procedure.

DISCUSSION

Clinical unmet needs

Clinical unmet needs related to OR scheduling can be identified based on 4 aspects: hospital management, surgical supporting staffs, patients and surgeons [24]. For the perspective of hospital management, it is difficult to strike a balance between decreasing idle time and avoiding overtime of staff to maximize OR utility as well as to reduce costs. Idle time is harmful to maintaining a cost effective OR because time available for a surgical procedure to be performed is not being used. Meanwhile, OR over-utilization is 2.5 times more costly than under-utilization [25]. For surgical supporting staffs, they have to work unexpected overtime and under stressful circumstances when actual durations are significantly longer than scheduled durations. This leads to job dissatisfaction and higher turnover rate, which may subsequently affect the quality of care and safety provided by supporting staffs as they are overloaded or overworked. From the perspective of patients, the uncertainty of not knowing when is the start time of their surgeries can introduce additional stress and may lead to significant patient dissatisfaction. As for younger surgeons, who are not assigned to have the first case, are more mindful of scheduling accuracy. They may need to wait on a prior case to be finished by a different surgeon when actual durations are longer than scheduled durations. Continued delays in the same OR may also result in lack of staff to cover late cases.

Insights and applications

Owing to the above mentioned needs, accurate prediction of surgical case duration plays a vital role in increasing OR efficiency, reducing costs, maintaining hospital reputation, as well as improving patient and surgeon satisfaction. This study not only helps to improve the accuracy of OR case prediction but also provides meaningful insights on how predictions were made by the developed ML model. It has both clinical and technical novelties in the following aspects. For the clinical perspective we modeled OR events as dependent events instead of independent. We extracted some additional information from surgeon data, e.g., previous working time and no. of previous surgeries of the primary surgeons within the last 7 days and 24 hr, and this information was taken into consideration during model building. Furthermore, while other past studies reported how they developed and evaluated their predictive models [7,15,16,26,27], global and local interpretability on model prediction were conducted in this study. Interpretation on model output unraveled those most surgical cases using anesthesia type of GA took longer time to be completed. GA was consistently revealed to be the most important feature contributing to longer durations in SHAP analysis of XGB, XGB1 and XGB2 model output. However, Figure 6C showed that GA cases from the Orthopedics specialty were shorter than GA cases from other specialties. When focusing on cases within the Orthopedics specialty, SHAP interaction value revealed that GA cases were indeed shorter than cases that used spinal anesthesia (results not shown). This is because GA was mostly used in upper-limb surgeries while spinal anesthesia was mostly used in lower-limb surgeries in Orthopedics specialty. Lower-limb surgeries can be quite challenging and usually consumes longer durations. On the other hand, brain or spinal surgery using microscope were shorter for cases with 3-6 procedure numbers. Spinal fusion with spinal instrumentation ≤ 6 segments was associated with longer durations in most cases.

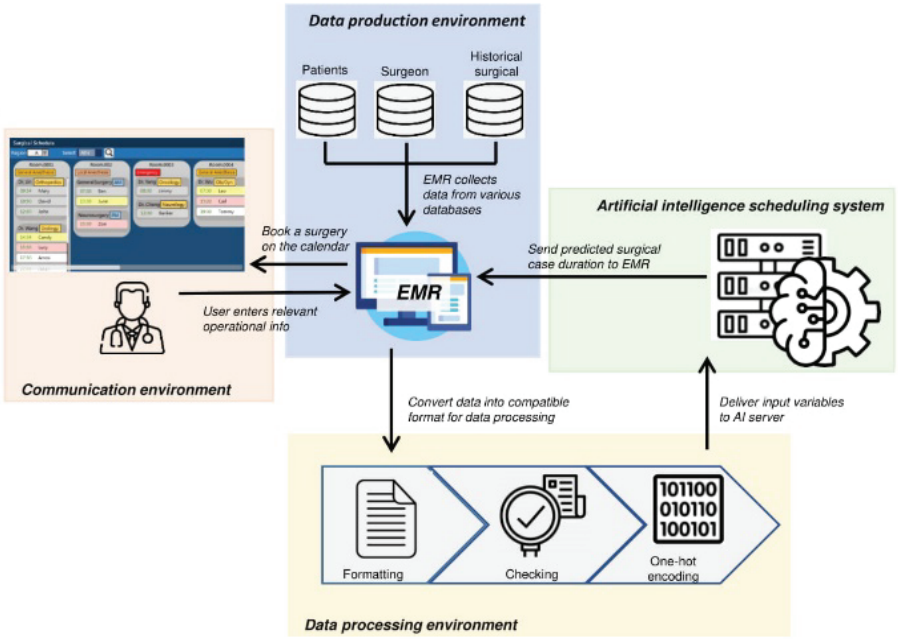

Figure 7: The mode of deployment of the machine learning (ML) models developed in this study. The ML models (XGB and XGB2) are deployed as a stand-alone machine intelligence (MI) server connected by the electrical medical record (EMR) system of the hospital. Upon receiving operational info in the EMR system, the EMR system collects data from various databases and converts them into compatible format for data processing. Subsequently, input variables are delivered to the MI server, where a prediction of surgical case duration is made by the XGB (for cases with 1 procedure) or XGB2 (for cases with ≥ 2 procedures) model. Subsequently, the MI server sends the output back to the EMR as a suggestion for the user to book a surgery on the calendar.

For the technical part, the data set used in this study contained approximately 80,000 cases (after exclusion) and more than 500 different types of surgical procedures, establishing a new benchmark for a massive quantity of data with high diversity. The maximal number of cases that had been used in other studies was in the range of 40,000 to 60,000 [7,16]. By using a large data set with huge diversity and variability, more powerful ML models could be developed. Interpretation on the ML model built with such a large and diverse data set subsequently aids in providing various clinically related information, which may not be disclosed by a simple ML model and is not discussed in the past studies [7,15,16,27]. Moreover, the developed XGB and XGB2 models are deployed as stand-alone machine intelligence (MI) server connected by the EMR of the hospital (Figure 7). Upon receiving relevant operational info, the EMR system collects data from various databases and converts them into compatible format for data processing. Subsequently, input variables are delivered to the MI server, where a prediction of surgical case duration is made by the XGB (for cases with 1 procedure) or XGB2 (for cases with ≥ 2 procedures) model. The MI server then sends the output back to the EMR as a suggestion for the user to book a surgery on the calendar. As the XGB1 model was not observed to improve predictive performance, the original XGB model is thus used to predict durations for cases with 1 procedure. All selected cases were used to train the XGB model while ~55% of the cases was used to build the XGB1 model and ~45% of the cases for the XGB2 model. For cases with 1 procedure, reducing training data size might have sacrificed some variations in the data that could be learned by the ML algorithm. On the contrary, cases with ≥ 2 procedures are relatively more complicated and contain more variations than cases with 1 procedure. Hence, removing cases with 1 procedure from training data did not sacrifice much data variation but indeed allows ML algorithm to focus on learning the various details in cases with ≥ 2 procedures.

Currently, surgical cases at CMUH are scheduled according to estimates made by primary surgeons, which were modeled as the averages for procedure in this study. However, many factors beyond expectation will not be taken into consideration. The performance of the average procedure-specific model, as reported in Figure 3, clearly showed that the predictions had larger and more scattered errors. When 20 feature variables (Table 3) were included in model development, the R^2, MAE, and percentages of underage, overage and within were improved substantially. When determining the tolerance threshold for percentage within, we set the criterion to be absolute duration differences falling within 0.15 x prediction or 30 minutes. Thirteen-minute difference in duration is usually considered as an acceptable periodic range for accurate booking. However, 30 minutes may be an excessively stringent standard for more complex and longer surgeries. Hence, 0.15 x prediction was also applied in the evaluation because a 15% error in prediction can typically be adapted by the operational management [15].

It has been reported in the past studies that primary surgeons contributed the largest variability in surgical case duration prediction compared to other factors attributed to patients [15,16,26]. These studies provide evidence and rationale that more factors relating to primary surgeons should be added as input variables in the development of ML models. Moreover, extensive feature engineering usually improves the quality of ML models and can be independent of the modeling technique itself. As a result, in addition to the primary surgeon's age and year of practice, we computed previous working time and number of previous surgeries performed by the same primary surgeons within the last 7 days and 24 hr. These variables extracted from primary surgeon data were significantly (p-value < 0.05) correlated with surgical case duration based on a regression analysis. The SHAP value distribution (Figure 5B) and SHAP interaction values (Figure 6D) of these features suggested that surgeons with moderate previous surgical time and more no. of surgical cases within a week took shorter time to complete the surgery. A practice effect was observed in these surgeons in whom they conducted similar surgeries multiple times and were able to finish the surgery in shorter duration. Whereas surgeons who had similar previous surgical time but lower no. of previous surgical cases in a week took longer time to complete the surgery.

While some studies included surgeon identity and operating location as input variables in their models [15,28], these were not included in feature inputs in this study. By doing so, our models can be generalized to new surgeons or new operating location in the hospital. Moreover, our models may be able to applied to other hospitals. However, there is no external data from other hospitals at the moment to verify the generalization of our models. There may be a need to fine-tune the model to better fit the settings of new environments or update our models after a while to fit the changes in medical technology. In terms of timing, we recommend updating the models annually by using surgical cases performed in the most recent 3 years as training data.

LIMITATIONS

One limitation in this study is that we selected predictor variables that could only be extracted from preoperative data. The scheduling system needs to be improved in order to be able to predict surgical case duration dynamically. For example, blood loss during surgery may affect case duration since an unexpected increase in blood loss may cause surgeons to take a longer time to complete the surgery. Therefore, it would be better if intra-operative data are incorporated during ML model development, and the prediction made by the ML model can be updated during surgery.

Meanwhile, one main reason that we only selected features which can be obtained pre-operatively is because the goal of building a predictive model is to improve and automate surgery scheduling before surgeries The model, however, does not serve to affect or restrict surgeons on how much time they would need to complete the surgery. Furthermore, one common issue in all ML studies in terms of predicting surgical case duration, including our study, is that ML models were developed using data from a single center. These ML models may have limitations in generalization since surgical team, facilities and patient populations are different across entities. A custom-made model has to be built for given and patient populations are different across entities. A custom-made model has to be built for a given facility itself. The other interesting issue of applying ML in surgical duration estimation is that medical technologies quickly evolve. Hence, how frequently an ML model needs to be updated still remains to be answered.

CONCLUSION

The XGB model was superior in predictive performance when compared to the average, Reg and RF models. Another ML

model, XGB2, was able to reduce prediction errors for cases with ≥ 2 procedures. We validated all models using a time-wise testing set in addition to the internal validation procedures. We validated all models using a time-wise testing set in addition to the internal validation generalized well to the time-wise testing data set even during the COVID-19 pandemic period. When external evaluation is not feasible, time-wise evaluation serves as a useful tool to better validate the predictive power of ML models. In addition to model development and evaluation, global and local interpretation on model output was conducted. Global interpretation aided in identifying features, e.g. anesthesia and procedure types that have important contribution and impact to surgical case durations while local interpretation revealed unique feature effects for a specific division or a subset of cases with specific conditions.

AUTHOR CONTRIBUTIONS

Conception and design: JL. and JY; Administrative support: JY; Provision of study materials or patients: DC; Collection and assembly of data: JH. and SL; Data analysis andinterpretation: JL; Manuscript writing: All authors; Final approval of manuscript: All authors

ACKNOWLEDGMENTS

The authors would like to thank Min-Hsuan Lu and the department of administration and management in CMUH in the assistance of data collection.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

FUNDING

This research received no external funding.S

REFERENCES

1. Gordon T, Paul S, Lyles A, Fountain J. Surgical unit time utilization review: Resource utilization and management implications. J Med Syst. 1988; 12(3):169-179.
2. Barbagallo S, Corradi L, De Goyet JD, Iannucci M, Porro I, Rosso N, et al: Optimization and planning of operating theatre activities: An original definition of pathways and process modeling. BMC Medical Inform Decis Mak. 2015; 15(1):1-6.
3. Childers CP, Maggard-Gibbons M. Understanding costs of care in the operating room. JAMA Surgery. 2018; 153(4):1-8.
4. Gillespie BM, Chaboyer W, Fairweather N. Factors that influence the expected length of operation: Results of a prospective study. BMJ Qual Saf. 2012; 21(1):3–12.
5. Levine WC, Dunn PF. Optimizing Operating Room Scheduling. Anesthesiol Clin. 2015; 33(4): 697–711.
6. Rothstein DH, Raval MV. Operating room efficiency. semin pediatr surg. 2018; 27(2):79–85.
7. Shahabikargar Z, Khanna S, Sattar A, Lind J. Improved Prediction of Procedure Duration for Elective Surgery. Stud Health Technol Inform. 2017; 239:133–138.
8. Laskin DM, Abubaker AO, Strauss RA. Accuracy of predicting the duration of a surgical operation. Journal of Oral and Maxillofacial Surgery 2013; 71: 446–447
9. May JH, Spangler WE, Strum DP, Vargas LG. The Surgical Scheduling Problem: Current Research and Future Opportunities. Production and Operations Management 2011; 20: 392–405
10. 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(1):41–49.
11. Zhou J,Dexter F, MacArio A, Lubarsky DA. Relying solely on historical surgical times to estimate accurately future surgical times is unlikely to reduce the average length of time cases finish late. Journal of Clinical Anesthesia. 1999; 11(7):601–605.
12. Kougias P, Tiwari V, Berger DH. Use of simulation to assess a statistically driven surgical scheduling system. Journal of Surgical Research. 2016; 201(2):306–312.
13. Hosseini N, Sir MY, Jankowski CJ, Pasupathy KS. Surgical Duration Estimation via Data Mining and Predictive Modeling: A Case Study. AMIA Annual Symposium proceedings. 2015; 478-479.
14. Edelman ER,van Kuijk SM, Hamaekers AE, de Korte MJ, van Merode GG, Buhre WF. Improving the prediction of total surgical procedure time using linear regression modeling. Frontiers in Medicine. 2017; 4(85):1-5.
15. Master N, Zhou Z, Miller D, Scheinker D, Bambos N, Glynn P. Improving predictions of pediatric surgical durations with supervised learning. International Journal of Data Science and Analytics. 2017; 4: 35–52.
16. Bartek MA, Saxena RC, Solomon S, Fong CT, Behara LD, Venigandla R.et al: Improving Operating Room Efficiency: Machine Learning Approach to Predict Case-Time Duration .J Am Coll Surg. 2019; 229(4):346–54.
17. Prasad AM, Iverson LR, Liaw A. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems. 2006; 9(2):181–199.
18. Nielsen D. Tree Boosting With XGBoost - Why Does XGBoost Win "Every" Machine Learning Competition? 2016.
19. Bentéjac C. Csörg oA,Martínez-Muñoz G. A Comparative Analysis of XG Boost. ArXiv. 2019.
20. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001.
21. Lundberg,SM, Lee SI. A Unified Approach to Interpreting Model Predictions.2017.
22. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et all. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020; 2(1): 56–67.
23. Ye K. The START methodology to work up and manage surgery scheduling inaccuracy. Technical report. 2016.
24. Sikka P. Basic Clinical Anesthesia. Springer. New York. United States. 2015.
25. Strum DP,Sampson AR, May JH, Vargas LG. Surgeon and type of anesthesia predict variability in surgical procedure times. Anesthesiology. 2000; 92(5):1454–1466.
26. Zhao B, Waterman RS, Urman RD, Gabriel RA. A Machine Learning Approach to Predicting Case Duration for Robot-Assisted Surgery. Journal of Medical Systems. 2019; 43(2):32.
27. Jiao Y, Sharma A, Ben Abdallah A, Maddox TM, Kannampallil T. Probabilistic forecasting of surgical case duration using machine learning: model development and validation. Journal of the American Medical Informatics Association. 2020; 27(12): 1885–1893.