

# Improving and Interpreting Surgical Case Duration Prediction with Machine Learning Methodology.

Jesyin Lai<sup>1,+</sup>, Ching-Chieh Huang<sup>1,+</sup>, Shu-Cheng Liu<sup>1</sup>, Jhao-Yu Huang<sup>1</sup>, Der-Yang Cho<sup>2</sup>, and Jiaxin Yu<sup>1,\*</sup>

<sup>1</sup>AI Innovation Center, China Medical University Hospital, Taichung, Taiwan

<sup>2</sup>Department of Neurosurgery, China Medical University Hospital, Taichung,  
Taiwan

## Abstract

Hospitals have encountered challenges in performing efficient scheduling and good resource management to ensure a high quality of healthcare is provided to their patients. Operating room (OR) scheduling is one of the issues that has gained our attention because it is related to workflow efficiency and critical care of hospitals. Automatic scheduling and high predictive accuracy of surgical case duration have a critical role in improving OR utilization. To estimate surgical case duration, most hospitals might rely on historical averages based on a specific surgeon or a specific procedure type obtained from electronic medical record (EMR) scheduling systems. However, the low predictive accuracy with EMR data leads to negative impacts on patients and hospitals, such as rescheduling of surgeries and cancellation. This study aims to improve and interpret the prediction of surgical case duration with machine learning (ML) methodology. A large data set containing 170,748 surgical cases (from Jan 2017 to Dec 2019) was obtained from a hospital, and it covered a broad variety of details on patients, surgeries, specialties and surgical teams. In addition, a more recent data set with 8,672 cases (from Mar to Apr 2020) was available to be used for time-wise evaluation. Historical averages were computed from the EMR data for surgeon- or procedure-specific cases, and served as baseline models for comparison. Subsequently, models were built with linear regression, random forest and extreme gradient boosting (XGB) algorithms, and were evaluated with R-square ( $R^2$ ), mean absolute error ( $MAE$ ), percentage overage (actual duration longer than prediction),

underage (shorter than prediction) and within (absolute duration differences falling within minimum(maximum(15 %,15 min), 60 min) of prediction). The XGB model was superior to the other models, achieving a higher  $R^2$  (84 %) as well as a lower  $MAE$  (31.1 min) and inaccurate percentage (24.4 %). In addition, XGB predictions were analyzed with Shapley additive explanations (SHAP). SHAP interpretation on complex cases (e.g. containing more than 2 procedures) unraveled that older primary surgeons took shorter time to complete the surgery and primary surgeons with longer previous surgical time within a week took more time to complete the surgery. Longer durations were utilized when patient's hypertension status was unknown. Meanwhile, SHAP interpretation on model loss showed that the loss values of elder primary surgeons increased for cases with larger deviations in prediction suggesting additional information related to surgeon is required for model improvement. Overall, this study applied ML techniques in the field of OR scheduling to reduce the medical and financial burden for healthcare management. The results revealed the impact of main factors (e.g. anesthesia, procedure types, no. of procedure) and interaction effects (e.g. no. of procedure x primary surgeon's age) in surgical case duration prediction as well as identified the feature that contributes to errors in prediction.

## 1 Introduction

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

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

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

21 Currently, most hospitals schedule surgical case duration by employing estimations from the  
22 surgeon and/or averages of historical case durations, and studies show that both of these methods  
23 have limited accuracy [8, 9]. For case lengths estimated by surgeons, factors including patient  
24 conditions and anesthetic issues might not be taken into consideration. Moreover, underestimation  
25 of case duration often occurs because surgeon estimations are usually made by favoring maximizing  
26 block scheduling to account for potential cancellations and cost reduction. Furthermore, operations  
27 with higher uncertainty and unexpected findings during surgery add difficulties and challenges to  
28 case length estimation [8]. Historical averages of case duration for a specific surgeon or a specific  
29 type of surgery obtained from electronic medical record (EMR) scheduling systems have also been  
30 used in hospitals. However, these methods have been shown to produce low accuracy due to the  
31 large variability and lack of the same combination of factors in the preoperative data available on  
32 the case that is being performed [10].

33 To improve the predictability, researchers have utilized linear statistical models, such as re-  
34 gression, or simulation for surgical duration prediction and evaluation of the importance of input  
35 variables [11, 12, 13]. However, a common shortcoming of these studies is that relatively fewer input  
36 variables or features were used in their models than in alternative approaches due to the limitation  
37 of statistical techniques in handling too many input variables. Recently, machine learning (ML) has  
38 been shown to be powerful and effective in aiding health care management. Master et al. (2017)  
39 trained multiple ML models, including decision tree regression, random forest regression, gradi-  
40 ent boosted regression trees and hybrid combinations, to automate prediction and classification of  
41 pediatric surgical durations [14]. Ensemble algorithms, implementing least-squares boosting and  
42 bagging models with ML, developed by Shahabikargar et al. (2017) were shown to reduce the error  
43 by 55 % compared to the original error [7]. With the use of a boosted regression tree, Zhao et al.  
44 (2019) increased the percentage of accurately booked cases for robot-assisted surgery from 35 %  
45 to 52 %. Bartek et al. (2019) reported that they were able to improve predicted cases within 10  
46 % of the threshold tolerance from 32 % to 39 % using an extreme gradient boosting model [15].  
47 Nonetheless, these ML studies included only 5-12 different types of procedures and specialties to

48 train their ML models, which may limit the generalization of these models.

49 In this study, more than 170,000 cases were obtained from China Medical University Hospital  
50 (CMUH) containing 743 types of procedures across 25 different specialties. From the original data,  
51 we further analyzed the working time of primary surgeons and computed their total number of  
52 previous surgeries and the total minutes spent on previous surgeries within 24 hr as well as within  
53 the last 7 days. Since surgeons' working performance might be affected by previous events, surgical  
54 cases performed by the same primary surgeon, especially within 24 hr, should not be considered  
55 as totally independent and unrelated. Hence, previous surgical counts and working time obtained  
56 from surgeons' data were included as additional features in our ML model training to account for  
57 their influences on surgical case duration. We hypothesize that these features impose significant  
58 influences on surgical case duration and may aid in improving the performance of a trained ML  
59 model.

## 60 Results

### 61 Model development and evaluation

62 Approximately 17 % of the cases were excluded from the original data from Jan 1, 2017, to Dec  
63 31, 2019, based on the exclusion criteria mentioned in Fig. S2 (Supplementary info). Therefore,  
64 142,445 cases containing more than 743 procedural categories and 25 specialties were included for  
65 predictive model development and evaluation. Furthermore, a recent data set collected from Mar  
66 1 to April 30, 2020, (7,231 cases after exclusion) was used in the time-wise evaluation to verify the  
67 robustness of the model in making predictions.

68 The results of all the metrics used to evaluate model performance on training, internal validation  
69 and time-wise testing sets are reported in Table S1 of the Supplementary info. Fig. 1 summarizes  
70 the results of the model evaluation on the time-wise testing set. The figure shows that the average  
71 model for the surgeon-specific scenario was not a good estimate for surgical case duration. The  
72 average model for the procedure-specific scenario had a lower percentage underage (actual duration  
73 shorter than prediction) and overage (longer than prediction) than the surgeon-specific average  
74 model. These differences were due to an extensive procedure classification in the procedure-specific  
75 model. However, the percentage underage was still quite high. Since no other information is taken  
76 into consideration in both of the average models except the duration of surgical cases that happened  
77 in the past, these models usually exhibit prediction bias and low accuracy.

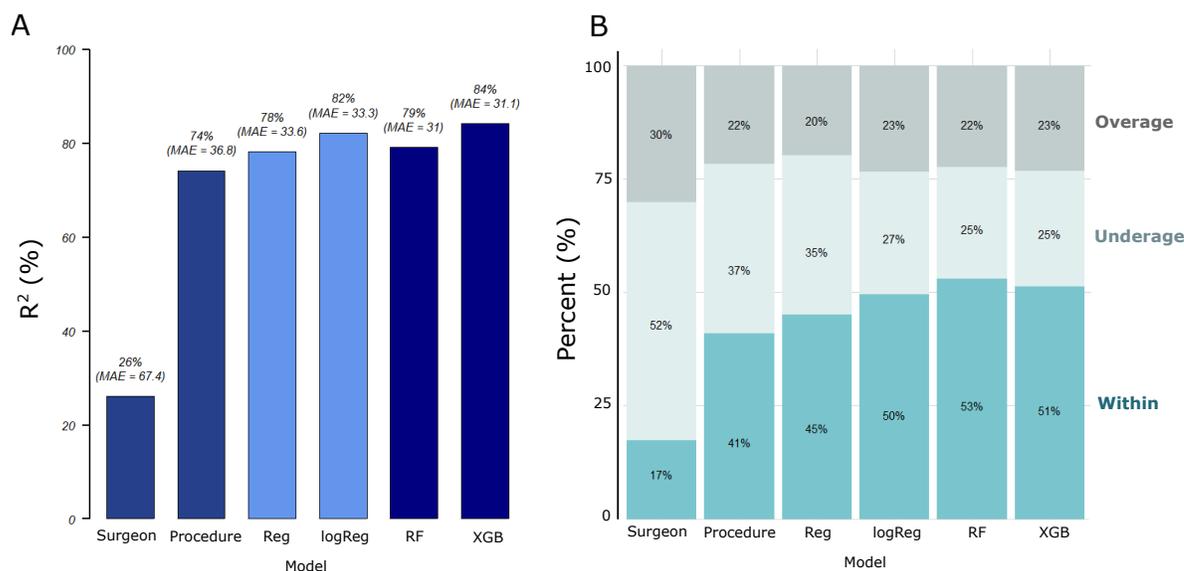


Figure 1: Machine learning algorithms improved the prediction accuracy of surgical case duration based on the results of time-wise model evaluation. The performance of all models was evaluated on a time-wise testing set (data not included in the original data set for ML model training) by using (A)  $R^2$ , mean absolute error ( $MAE$ ) as well as (B) percentage overage (actual duration longer than prediction), underage (shorter than prediction) and within (absolute duration differences falling within minimum(maximum(15 %,15 min), 60 min) of prediction).

Patient	Surgical team	Operation	Facility	Primary Surgeon's Prior Events
Age (age)	Surgeon team size (TeamSize)	Procedure type (Proced)	Day of the week (weekday)	No. of previous surgeries performed by the surgeon within 24 hr (Opcount_1d)
Gender (SexName)	Specialty (DivNo)	Anesthesia type (AnaValue)	Time of day (TimeofDay)	Total surgical minutes performed by the surgeon within 24 hr (Optoltime_1d)
ICD code (Diag)	Primary surgeon's age (Dr.age)	No. of procedure (Nproced)		No. of previous surgeries performed by the surgeon within the last 7 days (Opcount_7d)
In- /out-patient (OpType)				Total surgical minutes performed by the surgeon within the last 7 days (Optoltime_7d)
ASA status (ASA)				
Hypertension (Hypertension)				
BMI category (BMI)				

Table 1: Preoperative data with 19 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. The text in the 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 S3. ICD: International Classification of Diseases; ASA: American Society of Anesthesiologists; BMI: Body mass index.

78 We first fitted the linear regression (Reg) model by including all the input variables shown in  
79 Table 1. The evaluation metrics reported a higher percentage within than the average models on  
80 the time-wise testing sets (Fig. 1). There was an increase in the  $R^2$  value, indicating that predictive  
81 performance of the model improved when other information was taken into consideration. Since  
82 the distribution of the surgical case duration was found to be non-normal (Fig. S1A), we tested  
83 Box-Cox transformation on the Reg model. The results showed that the measure of goodness of  
84 fit was maximum when the optimal value for transformation was close to 0 (results not shown).  
85 Therefore, we applied log-transformation on the surgical case duration. The log-transformation is  
86 a special case of the Box-Cox transformation [16]. Moreover, by applying log-transformation, the  
87 differences between shorter durations were expanded while the differences between longer durations  
88 were reduced. This is because the slope of the logarithmic function is steeper for smaller values but  
89 moderate for larger values. When we log-transformed surgical case duration and reran a regression  
90 model (i.e., logReg), the performance of the logReg model outperformed the Reg model.

91 Although the performance of the logReg model was better than the Reg and the average models,  
92 an assumption of a linear relationship between the target and input variables was applied in both  
93 the Reg and the logReg models. The relationship between the target and input variables is usually  
94 nonlinear in a real-world situation. ML algorithms are helpful in making predictions in a more  
95 complicated scenario. The random forest (RF) model is the first ML model that we built in this  
96 study. There was a slight improvement in  $MAE$  compared to that of the logReg model. However,  
97 overfitting was observed in the RF model as performance of the training set was better than the  
98 internal validation and time-wise testing sets. An XGB model was subsequently developed to find  
99 out if the predictive performance would be better. Although the results of  $MAE$ , underage, within

	Actual	Surgeon-specific	Procedure-specific	Reg	logReg	RF	XGB
<b>Total minutes</b>	920,374	910,190	938,154	945,001	903,983	875,144	892,707
<b>Total prediction error in minutes</b>		487,357	266,185	243,233	240,980	224,476	224,643
<b>Inaccurate percentage (%)</b>		53	28.9	26.4	26.2	24.4	24.4

Table 2: The extreme gradient boosting (XGB) and random forest (RF) model produced the lowest percentage of cumulative inaccuracy among all the other models. Cumulative differences between actual and predicted case durations for all the models are shown in this table.

100 and overage were similar, the XGB model outperformed the RF model in terms of  $R^2$ . Hence, the  
101 XGB model was chosen as the best model and was used in subsequent analysis.

102 In addition to the three key metrics, we studied the accumulative inaccuracy of all the models  
103 by using the time-wise testing set. The total prediction error (in minutes) and the corresponding  
104 inaccurate percentage were calculated (Table 2). The actual total minutes represent the sum of  
105 surgical case durations for 7,231 cases in the time-wise testing set. The inaccurate percentage was  
106 derived from the percentage of total prediction error divided by the actual total minutes. The  
107 outcome shows that the inaccurate percentages of both the ML models (RF and XGB) were the  
108 lowest. Since the XGB model was not overfitted and had a higher  $R^2$  value, the XGB model was  
109 applied to improve the efficiency of OR scheduling. In addition to looking at the overall performance  
110 across all specialties, we compared the performance of the XGB model by further breaking it down  
111 to the specialty level on the time-wise testing set. Fig. S3 (Supplementary Info) shows the number  
112 of cases that were predicted as overage, underage and within for each specialty in the time-wise  
113 testing set. Most specialties had more cases that were predicted as within than as overage and  
114 underage. There were some exceptional specialties which had fewer cases of within because the  
115 total case numbers in these specialties were low. Moreover, the mean case durations were long for  
116 bariatric and metabolic as well as pediatric dentistry. As these cases were rare and required longer  
117 duration, this rendered difficulties for the model to predict accurately.

118 Subsequently, we plotted Bland-Altman (BA) plots using the time-wise testing set for the aver-  
119 age models (surgeon- and procedure-specific) and the XGB model (Fig. 2). The BA plots for these  
120 three models were plotted on y-axes in two different scales, one is (actual-prediction) (Fig. 2A-C)  
121 and the other is (actual-prediction)/prediction (Fig. 2D-F). Histograms for x- and y-axes were also  
122 plotted alongside the corresponding BA plots in Fig. 2A-C. The BA plots and the histograms for  
123 y-axes clearly show that predictions of the XGB model, especially the shorter durations, were closer  
124 to the actual durations compared to predictions of the two average models. When duration differ-  
125 ences were divided by predictions (Fig. 2D-F), they represent the deviations of differences from the

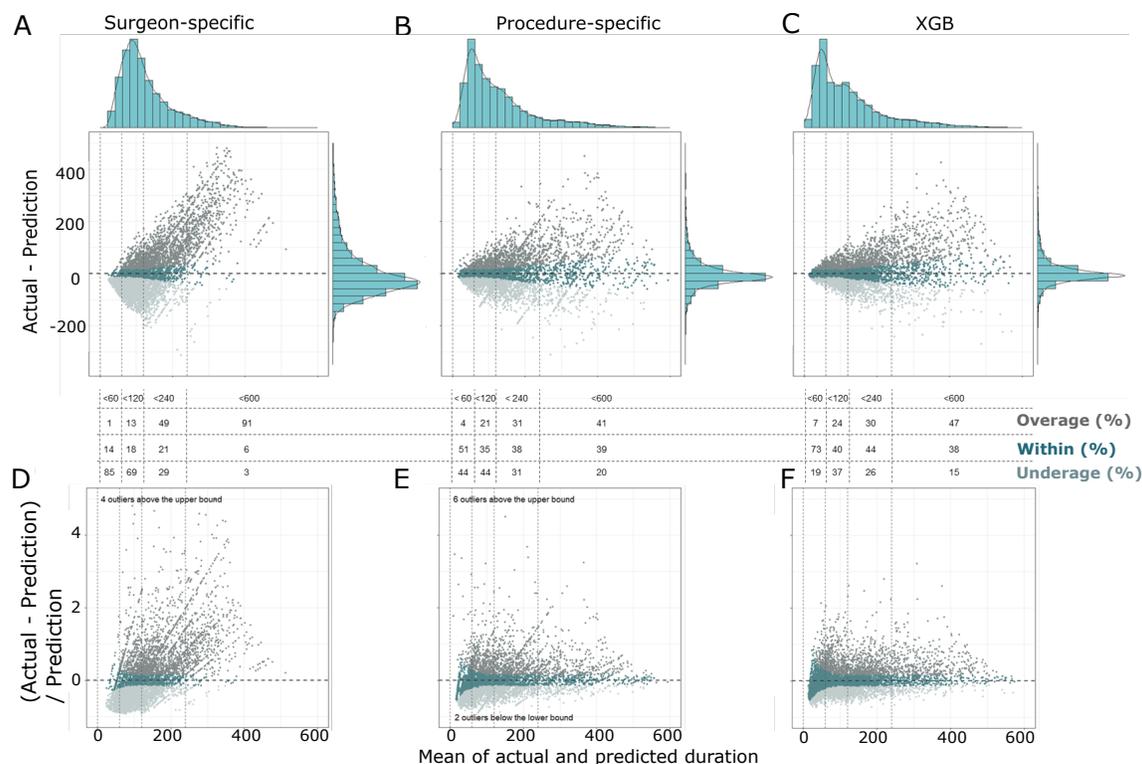


Figure 2: Predictions produced by the extreme gradient boosting (XGB) model, especially the shorter durations, were closer to the actual surgical case durations compared to predictions produced by average models. The Bland-Altman plots (with histograms for x- and y-axes) for average models of (A) surgeon- and (B) procedure-specific scenarios, and (C) the XGB model revealed that differences between actual and predicted durations in the XGB model were smaller. Bland-Altman plots with differences versus predicted durations for (D) surgeon- and (E) procedure-specific scenarios, as well as for (F) the XGB model illustrated that actual and predicted duration differences of the XGB model were smaller in deviations from the predictions. Percentage overage, within and underage for cases falling within 0-60 min, 60-120 min, 120-240 min and 240-600 min of these three models, respectively, were reported in the figure. Data points were colored according to the classification of overage, within and underage.

126 predictions. This helps quantifying to what extent the differences occurred relative to predictions.  
 127 In contrast to the average models, BA plot of the XGB model (Fig 2F) reveals deviations were  
 128 much smaller and less scattered, including cases that were considered as overage and underage as  
 129 well as long-duration cases. As a result, this demonstrates that the XGB model is more accurate  
 130 than the average models in predicting surgical case duration.

### 131 Model interpretation

132 To uncover the global importance and the impact of each feature on the XGB model output,  
 133 we extracted weighted feature gain (WFG) and applied Shapley additive explanations (SHAP) to

134 explain the model. WFG was computed based on the reduction in model accuracy when the variable  
135 was removed. This value serves as an indication of how important the variable is in improving the  
136 purity of a decision tree branch [17, 18]. A higher WFG percentage indicates that the variable  
137 is more important. Although WFG is widely used as the basis for feature selection methods, it  
138 may not be consistent with other feature attribution methods, such as split count or permutation  
139 [19]. In contrast, SHAP adopts the classical Shapley values estimation methods, which satisfy  
140 the desirable properties of local accuracy, missingness and consistency [20, 19]. It explains model  
141 output by computing the contribution of each feature to the prediction. Fig. 3A reports feature  
142 importance based on WFG while Fig. 3B shows feature impact on model output based on SHAP  
143 value. Both measures reveal that anesthesia type, procedure type, no. of procedure, hypertension  
144 and specialty were the 5 most important features in the XGB model. Notably, 3 of the top 5  
145 important variables were attributed to operative information (i.e., anesthesia type, procedure type  
146 and no. of procedure). Moreover, two of the features that we computed from the surgeon data  
147 (i.e., total surgical minutes and the number of previous surgeries performed by the surgeon within  
148 the last 7 days) were included within the top 10 list of both measures. Although there were some  
149 slight differences in the ranking order of feature importance, SHAP value is more informative as  
150 it provides quantification and visualization of the impact (negative or positive) of each feature on  
151 the final output for each case as well as the variations in feature contribution relative to changes in  
152 feature value (Fig. 4A).

153 While SHAP summary plot provides global interpretability reflecting the general behavior of the  
154 features in the model, we also investigated the specific behavior of the features in a subset of model  
155 predictions. As complex cases with more procedure numbers and larger team size may benefit more  
156 from this ML predictive approach, we extracted complex cases from the time-wise testing set for  
157 local interpretability. From the SHAP summary plot of one-hot encoded features in Fig. 4A, we  
158 observed that cases with anesthesia type of general anesthesia (GA) and out-patient cases were  
159 associated with higher model output (longer duration) in addition to larger no. of procedure and  
160 surgeon team size. In contrast, cases belong to the specialties of otolaryngology, head, and neck  
161 ('DivNo\_350'), and ophthalmology ('DivNo\_380') were associated with lower model output (shorter  
162 duration). Therefore, we selected complex cases by setting the criteria to be no. of procedure  $\geq$   
163 2, team size  $\geq$  2, anesthesia type = GA, and not in the specialties of otolaryngology, head, and  
164 neck, and ophthalmology. Fig. 4B is a heatmap of SHAP interaction values of one-hot encoded  
165 features for the extracted complex cases. Only one-hot encoded features with top 12 largest SHAP

166 interaction values are reported in the figure. Features with higher and interesting interactive effects  
167 were highlighted in red boxes and their SHAP interaction values were plotted in Fig. 4D-G. No.  
168 of procedure had the highest total SHAP interaction value among all other features. The plot of  
169 its SHAP main effect value (Fig. 4C) unravels that an increase in no. of procedure was highly  
170 associated with an increase in predicted duration. Features, including primary surgeon's age (Fig.  
171 4D), total surgical minutes performed by the surgeon within the last 7 days (Fig. 4E), and procedure  
172 type of spinal infusion surgery with spinal instrumentation equal or less than 6 segments ('83046',  
173 Fig. 4F), had higher SHAP interaction values with no. of procedure. From these SHAP interactive  
174 value plots, a few interesting phenomena were observed. For complex cases with 4-6 procedures,  
175 older primary surgeons were associated with shorter surgical case durations. On the other hand,  
176 primary surgeons with longer previous surgical minutes within the last 7 days were associated with  
177 longer durations for complex cases with 3-5 procedures. For procedure type of spinal infusion  
178 surgery with code '83046', surgical case durations were longer regardless of the no. of procedure.  
179 Total surgical minutes ('Optoltime\_7d') also interacted strongly with total no. ('Optolcount\_7d')  
180 of previous surgeries performed by the surgeon within the last 7 days (Fig. 4G). This reflects  
181 that primary surgeons with higher 'Optoltime\_7d' and higher 'Optolcount\_7d', but not moderate  
182 'Optoltime\_7d' and higher 'Optolcount\_7d', were associated with longer durations. In summary,  
183 these interactive effects suggest that older primary surgeons took shorter time to finish complex  
184 cases and primary surgeons with longer previous surgical minutes tended to take longer time to  
185 complete complex surgeries. Moreover, complex cases with procedure code '83046' consumed longer  
186 durations.

187 When we looked at SHAP interaction value for cases belong to the specialty of plastic and  
188 reconstruction ('DivNo\_411'), similar effect of primary surgeon's age was also observed. Fig. 4H  
189 shows that primary surgeons with younger age had larger positive impact on surgical case duration  
190 than those with older age implying that younger primary surgeons took longer time to perform  
191 the surgery. Moreover, for cases extracted from this specialty and the specialty of ophthalmology  
192 ('DivNo\_380'), cases with more no. of procedure and patient's hypertension status of unknown  
193 were associated with longer durations. However, this was not observed in cases with one procedure.  
194 These suggest that hypertension might be an important factor affecting surgical case duration.  
195 Hence, hypertension status of unknown might added in uncertainties and increased duration of  
196 cases with two or more procedure numbers.

197 In addition to global and local interpretability on model prediction, it is also important to

198 find out factors that contribute to errors in model prediction. This may subsequently help to  
199 debug and improve the model. SHAP loss value is a tool in SHAP package that can be used to  
200 find out each feature contribution to the model loss. Loss can be seen as a magnitude between  
201 the actual values and the predicted values. A larger loss implies a larger error in model output.  
202 Hence, lower model loss is preferable as this indicates that the predicted values are closer to the  
203 actual values. As Fig. 2C reveals that cases with shorter durations had less errors than longer  
204 durations in the XGB model, we extracted cases with predicted durations falling in the range of  
205 180-600 min from the testing set and investigated which features contribute to higher errors, i.e.  
206 higher loss. Subsequently, we selected cases with absolute duration differences less than 30 min  
207 (clinically not significant differences) or more than 90 min (large differences). Fig. 5A and B are  
208 summary plots of the top 10 one-hot encoded features with the highest average SHAP loss value  
209 magnitude for cases with small and large differences, respectively. The SHAP loss value of primary  
210 surgeon's age ('Dr\_age') was noticed because this feature had a longer positive tail, i.e. higher  
211 contribution to the model loss, in cases with large differences than cases with small differences.  
212 Meanwhile, other features had long negative tails implying that these features contributed to the  
213 reduction in model loss. We then focused on cases with predicted durations of 180-600 min to look  
214 at the distribution of SHAP loss value for one feature. We sorted all the cases according to their  
215 absolute differences normalized to predictions (which quantify the deviations of differences from  
216 the predictions) and plotted Fig. 5C, which manifests deviations (y-axis) were larger for cases with  
217 larger index numbers (x-axis). The SHAP loss value for the feature of primary surgeon's age in Fig.  
218 5D reveals a significant inconsistency (grey vertical line) in loss value as case index increases (i.e.,  
219 cases with larger deviations). The grey vertical dashed line indicates the location where the t-test  
220 (with Bonferroni correction) has a p-value less than  $0.05/(\text{no. of comparisons})$ . The t-test was  
221 automatically and repeated conducted to compare two data samples splitting with an increment of  
222 50 data points when plotting the figure with a function in the SHAP python package. When we  
223 manually compared SHAP loss values of the first 400 and last 400 cases with primary surgeon's  
224 age older than 65 years (Fig. 5D) using a one-sided t-test, the SHAP loss values of the latter were  
225 significantly ( $t = -2.64$ ,  $p\text{-value} = 0.01$ ) higher. These results suggest that primary surgeon's age is  
226 one of the factors contributing to the increase in prediction error. In contrast, the SHAP loss value  
227 for the feature of no. of procedure in Fig. 5E were consistent for cases with smaller and larger  
228 deviations. Although the mean SHAP loss value magnitude of no. of procedure was the largest, its  
229 contribution on model loss was observed to be not associated with the increase in error made by

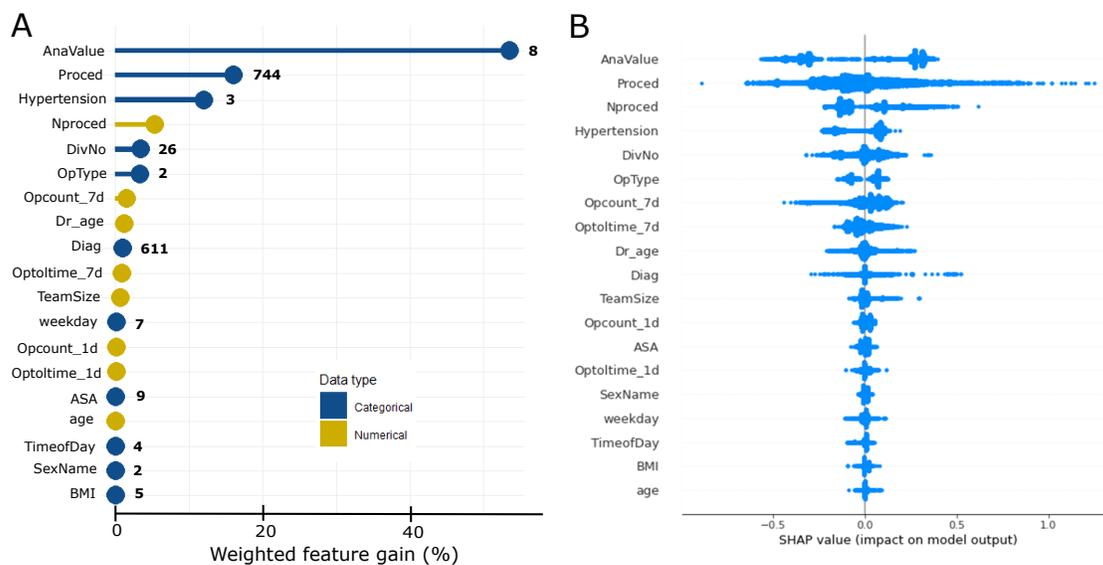


Figure 3: Anesthesia type, procedure type, number of procedure, hypertension and specialty were the top 5 important features used by the extreme gradient boosting (XGB) model to make predictions. Features were arranged and ranked in descending order according to (A) weighted feature gain or (B) SHAP value. In panel (A), categorical features are represented by blue markers, while numerical features are represented by yellow markers. The numbers of categories for categorical features are also shown beside the markers. The results of SHAP value show that total surgical minutes and number of previous surgeries performed by the surgeon within the last 7 days ('Opcount\_7d' and 'Optoltime\_7d') had important contributions to model output. Specific code names are used to represent the full names of features in the figure. The features' full names can be referred to Table 1.

230 the model.

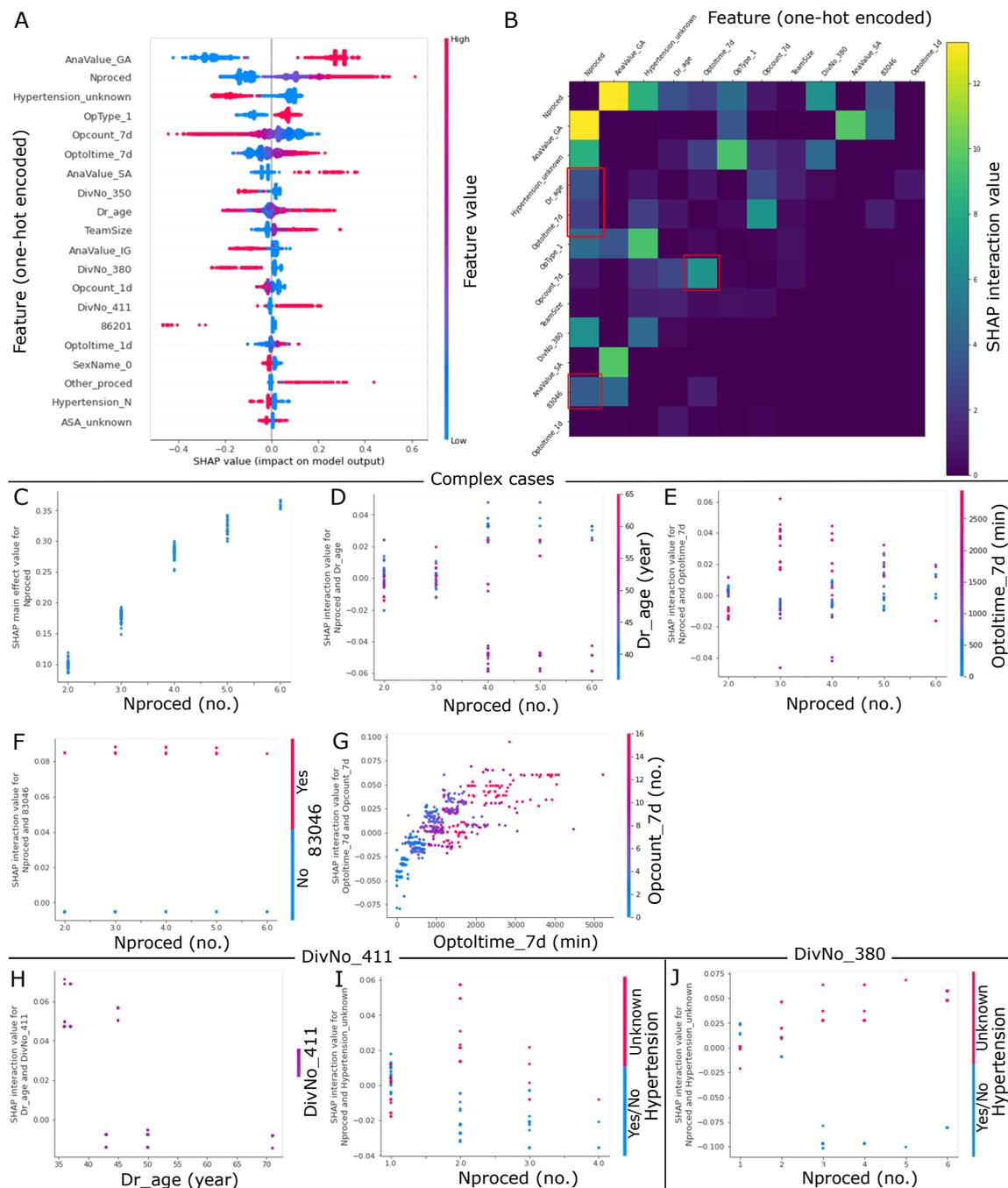


Figure 4: Shapley additive explanations (SHAP) interaction value unravels interactions between various one-hot encoded features. (A) A summary plot of 20 one-hot encoded features with the largest average SHAP value magnitude for the extreme gradient boosting (XGB) model. Each dot corresponds to a case in the time-wise testing set. No. of procedure ('Nproced'), anesthesia type of spinal anesthesia ('AnaValue\_SA'), Team size ('TeamSize'), specialty of plastic and reconstruction ('DivNo\_411'), and other procedure ('Other\_proced') had long positive tails. These indicate that these features tended to increase model output (longer surgical case duration) when feature values were either high (for numerical feature) or 1 (for categorical feature).

---

Figure 4 (*previous page*): (B) 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. (C) SHAP main effect value for no. procedure ('Nproced'). Larger no. of procedure had increased impact on model output. (D) SHAP interaction value for no. of procedure and primary surgeon's age ('Dr\_age'). For cases with 4-6 procedures, older primary surgeons were associated with shorter case durations. (E) SHAP interaction value for no. of procedure and total surgical minutes of previous surgeries performed by the surgeon within the last 7 days ('Optoltime\_7d'). For cases with 3-5 procedures, primary surgeons with larger 'Optoltime\_7d' were associated with longer case durations. (F) SHAP interaction value for no. of procedure and procedure type of spinal infusion surgery with spinal instrumentation equal or less than 6 segments ('83046'). Regardless of the no. of procedure, longer durations were consumed for cases with procedure code of '83046'. (G) SHAP interaction value for total surgical minutes ('Optoltime\_7d') and number ('Opcount\_7d') of previous surgeries performed by the surgeon within the last 7 days. The impact of 'Optiltime\_7d' on case durations was affected by changes in 'Opcount\_7d'. (H) and (I) were plotted for surgical cases belonged to the specialty of plastic and reconstruction ('DivNo\_411' = 1), while (J) was plotted for surgical cases belonged to the ophthalmology specialty ('DivNo\_380' = 1). (H) SHAP interaction value for primary surgeon's age and cases belong to the specialty of plastic and reconstruction. Younger primary surgeons took longer time to complete the surgery. (I) and (J) SHAP interaction value for no. of procedure and hypertension status. In both specialties of plastic and reconstruction, and ophthalmology, cases with hypertension status of unknown were associated with longer durations as no. of procedure increased.

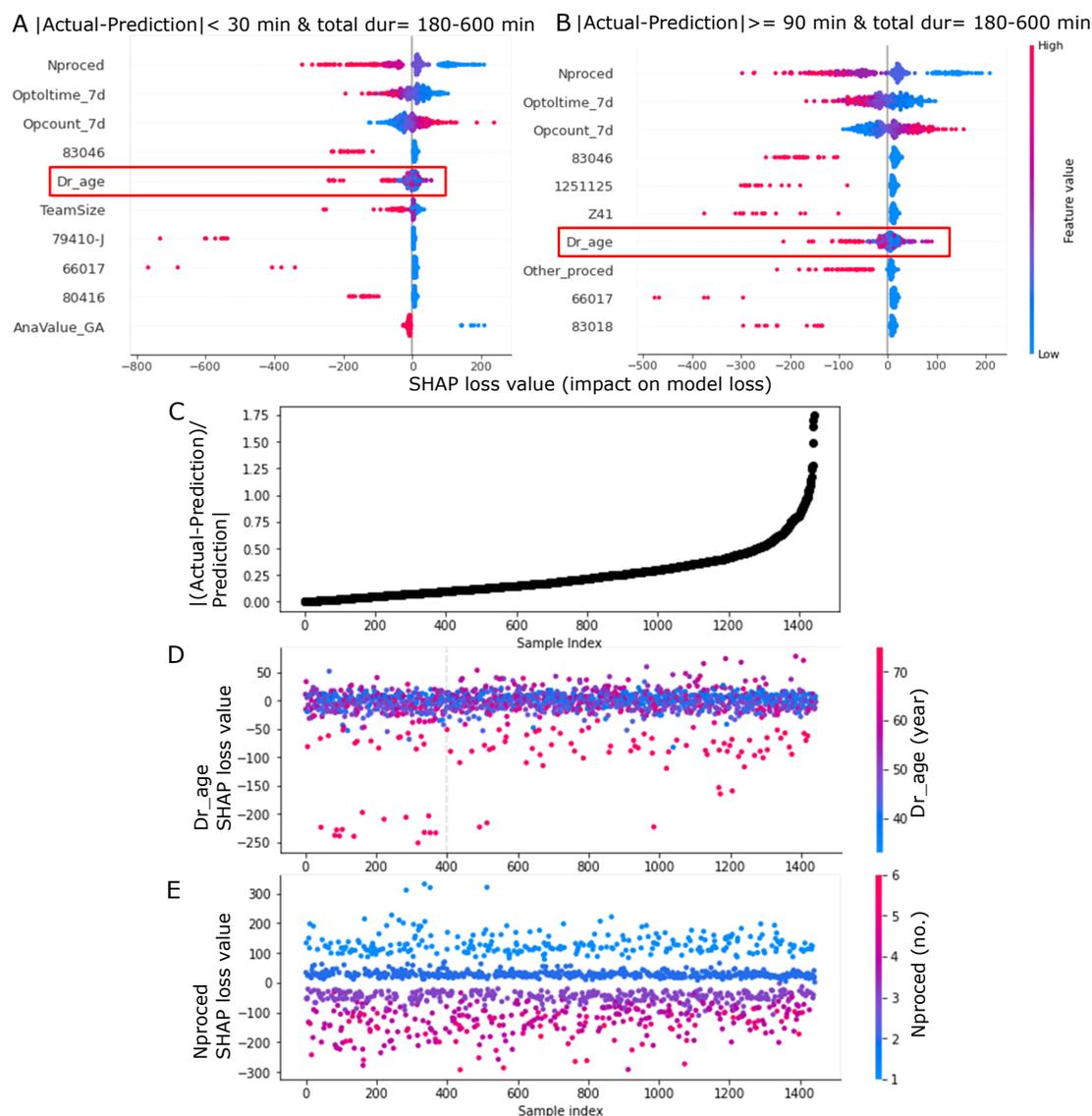


Figure 5: Shapley additive explanations (SHAP) loss value aids in identifying features with larger contribution to the model loss. Cases with predicted durations of 180-600 min were extracted from the time-wise testing set. Subsequently, cases with absolute duration differences ( $|\text{actual} - \text{prediction}|$ ) less than 30 min or more than 90 min were selected. Summary plots of 10 one-hot encoded features with the largest average SHAP loss value magnitude for (A) cases with differences less than 30 min (clinically not significant differences) and (B) cases with differences more than 90 min (large differences). Each dot corresponds to a case. The SHAP loss values of primary surgeon's age ('Dr\_age') are highlighted in red boxes in (A) and (B). This feature had a longer positive tail in cases with large differences compared to cases with small differences. (C) All cases with predicted durations of 180-600 min were sorted according to their absolute differences normalized to predictions ( $|\text{actual} - \text{prediction}|/\text{prediction}|$ ). The dot plot shows that deviations (y-axis) were larger for cases with larger index numbers (x-axis).

Figure 5 (*previous page*): (D) The SHAP loss value for the feature of primary surgeon's age. A grey vertical dashed line indicates the location where the t-test (with Bonferroni correction) has a p-value less than  $0.05/(\text{no. of comparisons})$ , i.e. this reflects the inconsistency of SHAP loss value for the feature. The t-test was automatically and repeated conducted to compare two data samples splitting with an increment of 50 data points when plotting the figure with a function in the SHAP python package. This indicates that primary surgeon's age is one of the factors contributing to the increase in prediction error. (E) The SHAP loss value for the feature of no. of procedure ('Nproced'). In contrast to primary surgeon's age, SHAP loss values for no. of procedure exhibited consistency for cases with smaller and larger deviations. Although the mean SHAP loss value magnitude of no. of procedure was the largest, it was not related to changes in error produced by the model.

## 231 Discussion

232 Clinical unmet needs related to OR scheduling can be identified based on the aspects of hospital  
233 management, surgical supporting staffs, patients and surgeons [21]. From the hospital manage-  
234 ment's point of view, it is difficult for them to strike a balance between decreasing idle time and  
235 avoiding overtime of staff to maximize OR utility as well as to reduce costs. Idle time is harmful  
236 to maintaining a cost effective OR because time available for a surgical procedure to be performed  
237 is not being used. Meanwhile, OR over-utilization is 2.5 times more costly than under-utilization  
238 [22]. For surgical supporting staffs, they have to work unexpected overtime and under stressful cir-  
239 cumstances when actual durations are significantly longer than scheduled surgical case durations.  
240 This may lead to job dissatisfaction and higher turnover rate, which may subsequently affect the  
241 quality of care and safety provided by supporting staffs as they are overload or overworked. From  
242 the perspective of patients, they often do not know when is the start time of their surgeries if the  
243 surgeries are not scheduled as the first case in the morning. Hence, prolonged wait times and in-  
244 creased uncertainties may lead to significant patient dissatisfaction. As for surgeons, new surgeons,  
245 who are not assigned to have the first case, are more mindful of scheduling accuracy. This is because  
246 they may need to wait on a prior case to be finished by a different surgeon when actual durations  
247 are longer than scheduled durations. Continued delays in the same OR may also result in lack of  
248 staff to cover late cases.

249 Owing to the above mentioned needs, accurate prediction of surgical case duration plays a vital  
250 role in increasing OR efficiency, reducing costs, maintaining hospital reputation, as well as improv-  
251 ing patient and surgeon satisfaction. This study not only helps to improve the accuracy of OR case  
252 prediction but also provides meaningful insights on how predictions were made by the developed  
253 ML model. It has both clinical and technical novelties in the following aspects. For the clinical

254 aspect, we are the first who modeled OR events as dependent events instead of independent. We  
255 extracted some additional information from surgeon data, e.g., previous working time and no. of  
256 previous surgeries of the primary surgeons within the last 7 days and 24 hr, and this information  
257 was taken into consideration during model building. Interpretation on model output unraveled  
258 that previous working time and no. of previous surgeries interacted with each other and produced  
259 differential effects on surgical case durations. For primary surgeons who performed surgeries con-  
260 taining more than 2 procedures, having longer cumulative previous surgeries time tended to affect  
261 their performance on the surgery (Fig. 4E G). Furthermore, while other past studies only reported  
262 how they developed and evaluated their predictive models [15, 23, 14, 24, 7], we went further to  
263 perform global and local interpretability on model prediction as well as identify factors contributing  
264 to errors in model prediction. To this end, older primary surgeons were observed to take shorter  
265 time to complete complex surgeries with 4-6 procedures and longer duration was required when  
266 patient's hypertension status was unknown. The phenomenon regarding patient's hypertension  
267 status was identified in more than one specialties suggesting knowing patient's hypertension status  
268 may be critical to surgical time management. Even though older primary surgeon might have more  
269 experiences and tended to finish surgeries faster (Fig. 6D H), loss values of primary surgeon's age  
270 older than 65 years were inconsistent compared to younger primary surgeons (Fig. 5D). Therefore,  
271 to improve model prediction, we suggest including more detailed information on years of experience  
272 for primary surgeons and the role of primary surgeons (e.g. executor or consultant) to be used as  
273 features for future ML model development.

274 For the technical aspect, the data set used in this study contained more than 140,000 cases (after  
275 exclusion) and more than 700 different types of surgical procedures, establishing a new benchmark  
276 for a massive quantity of data with high diversity. The maximal number of cases that had been used  
277 in other studies was in the range of 40,000 to 60,000 [15, 7]. By using a large data set with huge  
278 diversity and variability, a more powerful ML model was trained and built. Interpretation on the ML  
279 model built with such a large and diverse data set subsequently aids in providing various clinically  
280 related information, which may not be disclosed by a simple ML model and is not discussed in the  
281 past studies. Moreover, the developed XGB model in this study will be deployed as a stand-alone  
282 artificial intelligence (AI) server connected by the EMR of the hospital. When a user schedules  
283 a surgery using the EMR system, the EMR delivers input parameters to the AI server, where a  
284 prediction of surgical case duration is made, and then the AI server sends the output back to the  
285 EMR as a suggestion for the user (refer to Fig. S4 in Supplementary Info).

286 Currently, surgical cases at CMUH are scheduled according to estimates made by primary sur-  
287 geons. However, surgeon estimates rely heavily on prior experiences of the surgeons, and many  
288 factors beyond expectation will not be taken into consideration. Since there is no formal record on  
289 surgeon estimates, we used averages that were calculated based on a specific surgeon or procedure  
290 type on the time-wise testing set as our baseline models. The performance of these two average  
291 models, as reported in Fig. 1 and Table S1 (Supplementary info), clearly showed that they were  
292 poor in predicting surgical case duration. These models also tended to underpredict surgical case  
293 duration according to their BA plots in Figs. 2. When 19 feature variables (Table 1) were included  
294 in our model development, the  $R^2$ , MAE, and percentages of underage, overage and within im-  
295 proved substantially compared to the baseline models. When determining the tolerance threshold  
296 for percentage within, we set the criterion to be absolute duration differences falling within mini-  
297 mum(maximum(15 %,15 min), 60 min) of prediction. We applied 15 minutes because  $\pm 15$  minutes  
298 is an acceptable periodic range at CMUH to be considered as accurate booking. To avoid having  
299 an excessively stringent standard, we applied 15 % because a 15 % error in prediction can typically  
300 be adapted by the operational management [14]. For cases that took longer than 400 minutes, 15  
301 % of predicted duration could be more than 1 hour and so we adopted 60 minutes to ensure the  
302 threshold was not too lax.

303 By using regression and ML approaches, we were able to decrease the total prediction error  
304 (Table 2) of surgical case duration at CMUH. Among all the models, the performance of the XGB  
305 model was considered to be the best because it produced the highest  $R^2$  value and had the lowest  
306 inaccuracy. Moreover, the XGB model reduced the total prediction error (in minutes) to 224,643  
307 minutes. Since most ORs usually have multiple cases scheduled per day, the total prediction error  
308 represents the cumulative effect of total OR cases in the 2-month period of Mar to April 2020.  
309 This cumulative effect may eventually reflect a significant financial advantage in scheduling an  
310 additional operation case [25]. This approach would also lead to a significant cost reduction and  
311 increase in revenue because ORs are utilized appropriately and efficiently. When we evaluated the  
312 case numbers of overage, underage and within predicted by the XGB model at the specialty level,  
313 there were more case numbers falling within the acceptable thresholds for most of the specialties  
314 in the time-wise testing set (Fig. 1). This finding justifies that the performance of the XGB model  
315 can be generalized across specialties.

316 It has been reported in the past studies that primary surgeons contributed the largest variability  
317 in surgical case duration prediction compared to other factors attributed to patients [15, 23, 14].

318 These studies provide evidence and rationale that more factors relating to primary surgeons should  
319 be added as input variables in the training of ML models. Moreover, extensive feature engineering  
320 usually improves the quality of ML models and can be independent of the modeling technique  
321 itself. As a result, in addition to the primary surgeon's age, we computed previous working time  
322 and number of previous surgeries performed by the same primary surgeons within the last 7 days  
323 and 24 hr. These variables extracted from primary surgeon data were significantly ( $p < 0.05$ )  
324 correlated with surgical case duration (see Table S2 in the Supplementary info). The correlation  
325 coefficients and the impact distribution (Fig. 6A) of these variables also revealed that a surgical case  
326 duration of a primary surgeon may decrease as he or she becomes more familiar with the surgical  
327 procedure but may increase if his or her total surgical minutes are too long. Although performing a  
328 surgery multiple times on different patients may help a primary surgeon to be more efficient in his  
329 or her next operation, a long working time may also lead to lethargy and may affect the primary  
330 surgeon's performance. This explanation overlaps with our observations in Fig. 4E and F in which  
331 primary surgeons with longer previous surgical durations tended to take longer time to complete  
332 the surgery, especially for complex cases containing 3-5 procedures.

333 In the methodology of data processing, for predictor variables that contained many categories,  
334 we grouped categories that had less than 20 cases into a category named 'Others'. In addition to  
335 reducing the data dimensionality for categorical features, this grouping may aid in the generalization  
336 of our model, which implies that our model will still be able to predict the case duration even for  
337 operations that are rare. Moreover, our model may be applied to new surgeons and other hospitals  
338 since surgeon's identifier and room number were not included in feature inputs. However, there is  
339 still a need to update our model after a while or fine-tune the model to better fit the settings of  
340 other hospitals. In terms of timing, we recommend updating the model annually by using surgical  
341 cases performed in the most recent 3 years as training data.

342 One limitation in this study is that we selected predictor variables that could only be extracted  
343 from preoperative data. Our ML model still needs to be improved in order to be able to predict  
344 surgical case duration dynamically. For example, blood loss during surgery may affect case duration  
345 since an unexpected increase in blood loss may cause surgeons to take a longer time to complete  
346 the surgery. Therefore, it would be better if intra-operative data are incorporated during ML  
347 model development, and the prediction made by the ML model can be updated during surgery.  
348 Meanwhile, one main reason that we only selected features which can be obtained pre-operatively  
349 is because the goal of building a predictive model is to improve and automate surgery scheduling

350 before surgeries. The model, however, does not serve to affect or restrict surgeons on how much  
351 time they would need to complete the surgery. Furthermore, one common issue in all ML studies in  
352 terms of predicting surgical case duration, including our study, is that ML models were developed  
353 using data from a single site. These ML models have limitations in generalization since the surgical  
354 team, facilities and patient populations are different across entities. A custom-made model has to  
355 be built for a given organization using training data containing its patients, procedures, surgeons,  
356 medical staff, and the facility itself. As a result, the exact same ML model may not perform well  
357 when applied to another organization or hospital. The other interesting issue of applying ML or AI  
358 in surgical duration estimation is that medical technologies quickly evolve. Hence, how frequently  
359 an ML or AI model need to be updated still remains to be answered.

## 360 Conclusion

361 The XGB model was superior in predictive performance when compared to the average, Reg and  
362 logReg models. The total inaccurate percentage of the XGB model was the lowest among the  
363 other models. When compared to other ML studies, the XGB model built in this study also  
364 had a higher coefficient of determination ( $R^2$ ) and lower percentages of under- and overprediction  
365 [15, 24, 7]. Moreover, this model improves the current OR scheduling method at CMUH, which  
366 is based on estimates made by surgeons. We validated the model types using a time-wise testing  
367 set in addition to the internal validation set split from the original data used in model training.  
368 The results provide the evidence that the model generalized well to the time-wise testing data  
369 set even during the COVID-19 pandemic period. When external evaluation is not feasible, time-  
370 wise evaluation serves as a useful tool to better validate the predictive power of ML models. In  
371 addition to model development and evaluation, global and local interpretation on model output  
372 as well as identification of the source of errors in model prediction were conducted in this study.  
373 Based on the results, older primary surgeons took shorter time to complete the surgery but primary  
374 surgeons with longer previous surgical duration within the last 7 days took more time to complete  
375 the surgery. Longer duration was consumed when patient's hypertension status was unknown.  
376 Meanwhile, inconsistency in the loss value of primary surgeons aged above 65 years was found to  
377 be associated with increased deviations in case duration. Hence, we propose including additional  
378 information relevant to surgeon or surgical team composition, such as years of experience and the  
379 role of primary surgeon, to be used as predictor variables for future ML model training.

## 380 **Methods**

### 381 **Data sources.**

382 Data for this study were collected retrospectively from the EMR scheduling system of CMUH  
383 located in Taichung, Taiwan. The data set covered a broad variety of details about patients,  
384 surgeries, specialties and surgical teams. A total of 170,748 cases performed between Jan 1, 2017,  
385 and Dec 31, 2019, were used for model development. Additionally, 8,672 cases performed between  
386 Mar 1 and April 30, 2020, were used as data for time-wise model evaluation in this study. The  
387 proportion based on patient characteristics in the overall data set and time-wise testing set was  
388 reported in Table S3 in the Supplementary Info. Over 700 different types of procedures across  
389 25 surgical specialties were included in the training data set. Institutional review board approval  
390 (CMUH109-REC1-091) was obtained from CMUH before carrying out this study.

### 391 **Exclusion criteria, data processing and feature selection.**

392 Emergent and urgent surgical cases were removed since these two types of surgeries cannot be  
393 scheduled until they happen. Surgical case duration more than 10 hours or less than 10 minutes  
394 were also removed. Surgical records with missing values were excluded. Patients who were pregnant  
395 as well as duplicate cases were removed. The exclusion criteria are shown in Fig. S2. This approach  
396 resulted in a data set of 142,445 cases that were used for model training and validation. The same  
397 criteria were also applied to the data of Mar 1 to April 30, 2020, and 7,231 cases remained after  
398 exclusion.

399 Features were selected from available data sources based on literature review and discussion with  
400 surgeons and administrators of CMUH. Although the model performance could be enhanced by some  
401 postoperative information (e.g., total blood loss), these parameters cannot be used as features for  
402 model training because they were either missing or simply estimated by surgeons before surgery.  
403 Therefore, only variables that were available before surgery were selected for model development.  
404 Furthermore, the correlations of feature variables with the surgical case duration were checked by  
405 performing a regression analysis. Only those variables with significant ( $p$ -value  $< 0.05$ ) correlation  
406 coefficients were selected as predictor variables for model training.

407 When visualizing all the categories of procedure type and the International Classification of  
408 Diseases (ICD) code, there were hundreds to thousands of categories in these two variables. To  
409 reduce the problem of having too many dimensions during one-hot encoding of categorical features,

410 we combined categories that had less than 20 cases in the training set into a category and named  
411 it as ‘Others’. Similarly, we combined categories for specialty that had less than 20 cases into the  
412 category of ‘Others’.

413 In addition, since surgical case duration can be related to the performance of surgeons and  
414 surgeons’ performance is affected by their working time, we analyzed primary surgeons’ previous  
415 surgical events. The number of previous surgeries and total surgical minutes performed by the same  
416 primary surgeons within the last 7 days and 24 hr. Together, 19 predictor variables were included  
417 for predictive model building in this study. These predictors can be categorized into 5 groups:  
418 patient, surgical team, operation, facility and primary surgeon’s prior events (see Table 1).

## 419 **Model development and training.**

420 We applied multiple ML methods for surgical case duration prediction. Surgical case duration (in  
421 minutes) is the total period starting from the time the patient enters the OR to the time of exiting  
422 the OR. Historical averages of case durations based on surgeon-specific or procedure-specific data  
423 from EMR systems were used as baseline models for comparison in case duration prediction. At  
424 the beginning, we performed multivariate linear regression (Reg) to predict surgical case duration.  
425 However, when we evaluated the distribution of surgical case duration, it was observed to be skewed  
426 to the right (Fig. S1 in the Supplementary info). We performed a logarithmic transformation on the  
427 surgical case duration to reduce the skewness. The model built from log-transformed multivariate  
428 linear regression (logReg) outperformed Reg in all evaluation indexes. Subsequent ML algorithms  
429 were also trained by using the log-transformed case duration as the target.

430 The first ML algorithm that we tested was random forest (RF), a tree-based supervised learning  
431 algorithm. RF uses bootstrap aggregation or a bagging technique for regression by constructing a  
432 multitude of decision trees based on training data and outputting the mean predicted value from  
433 the individual trees [26]. Tree-based techniques were suitable for our data since they include a large  
434 number of categorical variables, e.g., ICD code and procedure type, of which most were sparse.  
435 The extreme gradient boosting (XGB) algorithm is the other supervised ML algorithm that was  
436 tested for comparison to RF. Recently, the XGB algorithm has gained popularity within the data  
437 science community due to its ability in overcoming the curse of dimensionality as well as capturing  
438 the interaction of variables [27].

439 XGB is also a decision tree-based algorithm similar to the RF algorithm. However, the XGB  
440 and RF algorithms are different in the way that the trees are built. It has been shown that XGB

441 performs better than RF if parameters are tuned carefully; otherwise, it would be more likely to  
442 overfit if the data are noisy [28, 29]. For both RF and XGB algorithms, we adopted a 5-fold cross-  
443 validation strategy to tune the best hyperparameters, e.g. no. of estimators, maximum of depths,  
444 etc. For the XGB model, the best hyperparameters were number of estimators = 535,  $\eta = 0.5$   
445 (step size shrinkage to prevent overfitting), maximum of depths = 3,  $\gamma = 0.3$  (minimum loss  
446 reduction, where a larger  $\gamma$  represents a more conservative algorithm) and  $\alpha = 1$  (L1 regularization  
447 weighting term, where a larger value indicates a more conservative model). For the RF model, the  
448 best hyperparameters were number of estimators = 150, minimum samples split = 10, minimum  
449 samples leaf = 2 and maximum depth = 125.

450 A data-splitting strategy was used in the training for all the models. We randomly separated  
451 the data into training and validation subsets at a ratio of 4:1. The training data were used to  
452 build different predictive models as well as to extract important predictor variables. The validation  
453 data were used for internal evaluation of the models. In addition to interval evaluation, time-wise  
454 evaluation on all the models was performed using data from Mar 1 to Apr 30, 2020. These data  
455 were not included in the original data set for ML model training. The results obtained from time-  
456 wise evaluation are better in verifying the robustness of the trained model in making an accurate  
457 prediction since the data are temporally segregated from the original data. Historical averages of  
458 case duration for surgeon- or procedure-specific data calculated from EMR data were also evaluated  
459 on the same internal validation and time-wise testing sets to ensure fair and uniform comparison  
460 across all models. Data processing and cleaning as well as model development in this study were  
461 performed using R software. RF model development was performed using python.

## 462 **Model evaluation.**

463 Multiple predictive models were built to predict surgical case duration. Different standards are  
464 usually applied to evaluate the predictive performance of the built models. The three key metrics  
465 used to evaluate model performance in this study included (1) R-square ( $R^2$ ), (2) mean absolute  
466 error (MAE), and (3) the percentage overage, underage and within.

467  $R^2$  is the coefficient of determination; it represents the proportion of the variance for the actual  
468 case duration that is explained by predictor variables in our models.

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

469 MAE measures the average of errors between the actual case durations and the predictions.

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

470 In this study, percentage within indicates the percentage of cases with absolute duration dif-  
471 ferences falling within the threshold ( $\tau(\hat{y})$ ) of minimum(maximum(15 %,15 min), 60 min). The  
472 mathematical expression of case duration differences that satisfy the within condition is shown as  
473 below:

$$|y - \hat{y}| < \tau(\hat{y}) = \min \{ \max \{ 0.15 * \hat{y}, 15 \}, 60 \} \quad (3)$$

474 Meanwhile, percentage underage is the percentage of cases with actual case duration shorter  
475 than prediction and case duration difference was more negative than the threshold. Similarly,  
476 percentage overage is the percentage of cases with actual case duration longer than prediction and  
477 case duration difference was more positive than the threshold. The condition that defines a case as  
478 overage, within and underage is summarised as follows:

$$condition = \begin{bmatrix} overage, y - \hat{y} \geq \tau(\hat{y}) \\ within, |y - \hat{y}| < \tau(\hat{y}) \\ underage, y - \hat{y} \leq -\tau(\hat{y}) \end{bmatrix} \quad (4)$$

### 479 Feature importance based on weighted feature gain.

480 During the development of XGB model, feature importance type of ‘gain’ was applied. ‘Gain’ is  
481 the most relevant attribute to interpret the relative importance of each feature. In XGB algorithm,  
482 ‘gain’ is the average gain across all splits where feature was used and is determined by the improve-  
483 ment in accuracy brought by a feature to the branches it is on [30]. The calculation of ‘gain’ can  
484 be shown as below:

$$Gain = \frac{1}{2} \left[ \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (5)$$

485 In the above equation, G is the first derivative of training loss and H is the second derivative  
486 of training loss.  $\lambda$  represents the regularization rate while  $\gamma$  represents the minimum loss reduction  
487 required to make a split. Overall, this equation consists of the score on the new left (L) leaf, the  
488 score on the new right (R) leaf, the score on the original leaf, and regularization on the additional

489 leaf. Hence, if ‘gain’ is 0 or negative (i.e.  $\leq \gamma$ ), it indicates better not to add that branch. When  
490 analyzing feature importance of the model, the weighted feature gain was calculated by multiplying  
491 the percentage gain of each feature by their percentage frequency of occurrence in the model [15].

## 492 Model interpretation with SHAP

493 To further interpret the developed XGB model, we determined the Shapley additive explanations  
494 (SHAP) value by applying SHAP package in Python. SHAP can be used to explain the predicted  
495 output by computing the contribution of each feature to the prediction [31, 20]. SHAP value of a  
496 feature  $i$ , termed  $\phi_i$ , can be obtained with the following equation:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \left[ \frac{|S|!(|F| - |S| - 1)!}{|F|!} \right] [f(S \cup \{i\}) - f(S)] \quad (6)$$

497 In the above equation,  $F$  is the set of all features considered for the XGB algorithm,  $S$  denotes  
498 a subset of features obtained from the set  $F$  except feature  $i$ , and  $f(S)$  is the expected output given  
499 by the set  $S$  of features. In summary, SHAP values indicate the impact of a feature on the model  
500 output. For our ML model, a large positive (negative) SHAP value of a feature implies that this  
501 feature has a large contribution in predicting a longer (shorter) surgical case duration. Meanwhile,  
502 a SHAP value of 0 implies that this feature have no or low contribution in predicting surgical case  
503 duration. SHAP values are expressed in log-odds (7) in this study.

$$\log[P(\phi)/(1 - P(\phi))], P(\phi) < 1 \quad (7)$$

504 Local interaction effects between features were identified by applying SHAP interaction values  
505 in the SHAP package. While SHAP value is the attribution for each feature, SHAP interaction  
506 value is a matrix of feature attributions [32]. The interaction effects on the off-diagonal and the  
507 main effects are on the diagonal. The SHAP interaction values is defined as:

$$\Phi_{i,j}(f, x) = \sum_{S \subseteq M \setminus \{i,j\}} \frac{|S|!(M - |S| - 2)!}{2(M - 1)!} \nabla_{ij}(f, x, S) \quad (8)$$

508 when  $i \neq j$  and:

$$\nabla_{ij}(f, x, S) = f_x(S \cup \{i, j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S) \quad (9)$$

509 In the equation (8),  $m$  is the set of all  $M$  input features. More details regarding the computation  
510 of SHAP interaction values can be referred to Lundberg et al. (2020). The SHAP interaction value  
511 between feature  $i$  and feature  $j$  is split equally between each feature ( $\Phi_{i,j}(f, x) = \Phi_{j,i}(f, x)$ ), and  
512 the total interaction effect is the sum of  $\Phi_{i,j}(f, x)$  and  $\Phi_{j,i}(f, x)$ .

513 SHAP values of a model's loss function decompose the model loss into the loss for each model  
514 input feature [32]. They provide information on how each feature contributes to the model loss.  
515 The definition of the sum of SHAP loss value of all features is shown as below [33]:

$$\sum_{i=1}^N \text{Contrib}_{f_i} = \text{ModelLoss} - \text{Expected} \quad (10)$$

516 In the above equation, the calculation of expected loss is performed by first setting all the data  
517 labels to the average of all predictions, and then computes the average loss.  $\text{Contrib}_{f_i}$  represents  
518 the contribution of each feature to the model loss from the expected loss.

## 519 Data availability

520 The minimum data set (March to April 2020) used in time-wise evaluation for this study is available  
521 from our web site: <https://cmuhopai.azurewebsites.net/>. The data set required to replicate model  
522 training and internal evaluation contains personal data and is not publicly available, in keeping  
523 with the Data Protection Policy of CMUH.

## 524 Code availability

525 The code for data processing and model development in this study are presented on github:  
526 <https://github.com/AII-CMUH/AI-surgery-scheduling>.

## 527 References

528 [1] Gordon, T., Paul, S., Lyles, A. & Fountain, J. Surgical unit time utilization review: Resource  
529 utilization and management implications. *Journal of Medical Systems* **12**, 169–179 (1988).

- 530 [2] Barbagallo, S. *et al.* Optimization and planning of operating theatre activities: An original  
531 definition of pathways and process modeling. *BMC Medical Informatics and Decision Making*  
532 **15** (2015).
- 533 [3] Childers, C. P. & Maggard-Gibbons, M. Understanding costs of care in the operating room.  
534 *JAMA Surgery* **153** (2018).
- 535 [4] Gillespie, B. M., Chaboyer, W. & Fairweather, N. Factors that influence the expected length  
536 of operation: Results of a prospective study. *BMJ Quality and Safety* **21**, 3–12 (2012).
- 537 [5] Levine, W. C. & Dunn, P. F. Optimizing Operating Room Scheduling. *Anesthesiology Clinics*  
538 **33**, 697–711 (2015).
- 539 [6] Rothstein, D. H. & Raval, M. V. Operating room efficiency. *Seminars in Pediatric Surgery*  
540 **27**, 79–85 (2018).
- 541 [7] Shahabikargar, Z., Khanna, S., Sattar, A. & Lind, J. Improved Prediction of Procedure  
542 Duration for Elective Surgery. *Studies in health technology and informatics* **239**, 133–138  
543 (2017). URL <http://www.ncbi.nlm.nih.gov/pubmed/28756448>.
- 544 [8] Laskin, D. M., Abubaker, A. O. & Strauss, R. A. Accuracy of predicting the duration of a  
545 surgical operation. *Journal of Oral and Maxillofacial Surgery* **71**, 446–447 (2013).
- 546 [9] May, J. H., Spangler, W. E., Strum, D. P. & Vargas, L. G. The Surgical Scheduling Problem:  
547 Current Research and Future Opportunities. *Production and Operations Management* **20**,  
548 392–405 (2011). URL <http://doi.wiley.com/10.1111/j.1937-5956.2011.01221.x>.
- 549 [10] Zhou, J., Dexter, F., MacArio, A. & Lubarsky, D. A. Relying solely on historical surgical times  
550 to estimate accurately future surgical times is unlikely to reduce the average length of time  
551 cases finish late. *Journal of Clinical Anesthesia* **11**, 601–605 (1999).
- 552 [11] Kougias, P., Tiwari, V. & Berger, D. H. Use of simulation to assess a statistically driven  
553 surgical scheduling system. *Journal of Surgical Research* **201**, 306–312 (2016).
- 554 [12] Hosseini, N., Sir, M. Y., Jankowski, C. J. & Pasupathy, K. S. Surgical Duration Estimation  
555 via Data Mining and Predictive Modeling: A Case Study. In *AMIA Annual Symposium*  
556 *proceedings*, 640–8 (Mayo Clinic, 2015).

- 557 [13] Eijkemans, M. J. *et al.* Predicting the unpredictable: A new prediction model for operating  
558 room times using individual characteristics and the surgeon's estimate. *Anesthesiology* **112**,  
559 41–49 (2010).
- 560 [14] Master, N. *et al.* Improving predictions of pediatric surgical durations with supervised learning.  
561 *International Journal of Data Science and Analytics* **4**, 35–52 (2017).
- 562 [15] Bartek, M. A. *et al.* Improving Operating Room Efficiency: Machine Learning Approach to  
563 Predict Case-Time Duration. *Journal of the American College of Surgeons* **229**, 346–54 (2019).
- [16] Daimon, T. Box–Cox Transformation. In *International Encyclopedia  
of Statistical Science*, 176–178 (Springer Berlin Heidelberg, 2011). URL  
[https://link.springer.com/referenceworkentry/10.1007/978-3-642-04898-2\\_52](https://link.springer.com/referenceworkentry/10.1007/978-3-642-04898-2_52).
- 564 [17] Chen, T., He, T., Benesty, M. & Khotilovich, V. Package 'xgboost' Type Package Title Extreme  
565 Gradient Boosting (2020). URL <https://github.com/dmlc/xgboost/issues>.
- 566 [18] Song, Y. Y. & Lu, Y. Decision tree methods: applications for classification and prediction.  
567 *Shanghai Archives of Psychiatry* **27**, 130–135 (2015).
- 568 [19] Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent Individualized Feature Attribution for  
569 Tree Ensembles (2019). URL <http://arxiv.org/abs/1802.03888>.
- 570 [20] Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions (31st  
571 Conference on Neural Information Processing Systems, Long Beach, CA, 2017). URL  
572 <https://github.com/slundberg/shap>.
- 573 [21] Ye, K. The START methodology to work up and manage surgery scheduling inaccuracy. Tech.  
574 Rep. (2016).
- 575 [22] Sikka, P. *Basic Clinical Anesthesia* (Springer, 2015).
- 576 [23] Strum, D. P., Sampson, A. R., May, J. H. & Vargas, L. G. Surgeon and type of anesthesia  
577 predict variability in surgical procedure times. *Anesthesiology* **92**, 1454–1466 (2000).
- 578 [24] Zhao, B., Waterman, R. S., Urman, R. D. & Gabriel, R. A. A Machine Learning Approach  
579 to Predicting Case Duration for Robot-Assisted Surgery. *Journal of medical systems* **43**, 32  
580 (2019). URL <http://www.ncbi.nlm.nih.gov/pubmed/30612192>.

- 581 [25] Dexter, F. & Macario, A. Decrease in Case Duration Required to Complete an Additional  
582 Case During Regularly Scheduled Hours in an Operating Room Suite. *Anesthesia & Analgesia*  
583 **88**, 72–76 (1999). URL <http://journals.lww.com/00000539-199901000-00014>.
- 584 [26] Prasad, A. M., Iverson, L. R. & Liaw, A. Newer classification and regression tree techniques:  
585 Bagging and random forests for ecological prediction. *Ecosystems* **9**, 181–199 (2006).
- 586 [27] Nielsen, D. Tree Boosting With XGBoost - Why Does XGBoost Win "Every" Machine Learn-  
587 ing Competition? *Tree Boosting With XGBoost - Why Does XGBoost Win "Every" Machine*  
588 *Learning Competition?* (2016).
- 589 [28] Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A Comparative Analysis of XGBoost (2019).  
590 URL <http://arxiv.org/abs/1911.01914>.
- 591 [29] Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Annals of*  
592 *Statistics* (2001).
- 593 [30] Introduction to Boosted Trees — xgboost 1.3.0-SNAPSHOT documentation. URL  
594 <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>.
- 595 [31] 5.10 SHAP (SHapley Additive exPlanations) — Interpretable Machine Learning. URL  
596 <https://christophm.github.io/interpretable-ml-book/shap.html#shap>.
- 597 [32] Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI  
598 for trees. *Nature Machine Intelligence* **2**, 56–67 (2020).
- 599 [33] Lin, C. Use SHAP loss values to debug/monitor your model. URL  
600 <https://towardsdatascience.com/use-shap-loss-values-to-debug-monitor-your-model-83f7808af40f>.

## 601 Acknowledgments

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

## 604 Author contributions

605 + J.L. and C.H. contributed equally to this work. J.L. and C.H. performed the literature search,  
606 conducted the modelling, evaluation and model interpretation, contributed to the modelling design,

607 generated the figures and tables, and co-wrote the paper. S.L. and J.H built the website for data  
608 access, presented code on github, and assisted in model deployment. D.C. conceived the study and  
609 contributed to data interpretation. J.Y. contributed to the study design, data interpretation, and  
610 paper editing.

## 611 Competing interests

612 The authors declare that they have no competing interests.

## 613 Materials & correspondence

614 Correspondence and requests for materials should be addressed to Jiaxin Yu (email: [jiaxin.yu@mail.cmuh.org.tw](mailto:jiaxin.yu@mail.cmuh.org.tw)).

## 615 Supplementary Info

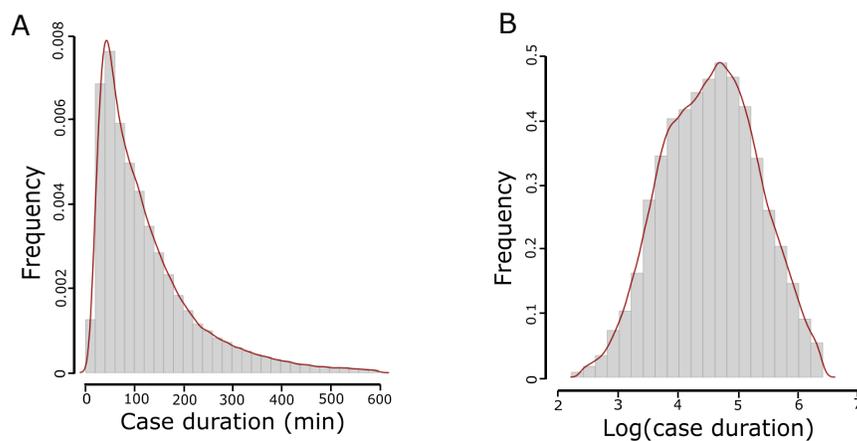


Figure S1: Log transformation of case duration converted the distribution of surgical case duration from (A) skewing to the right to (B) a more normal distribution.

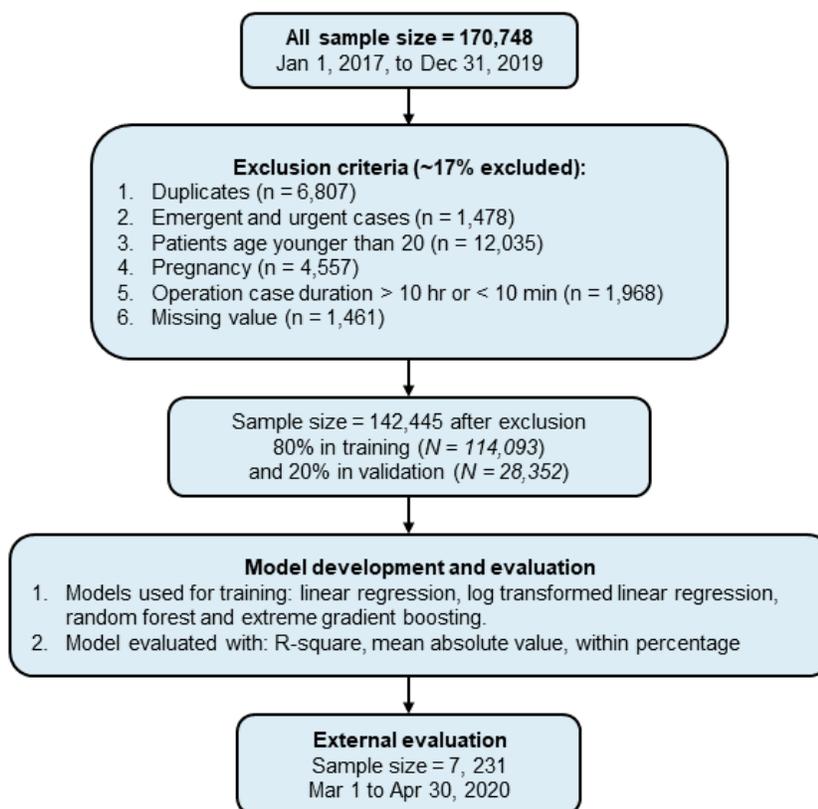


Figure S2: The workflow of model training for this study. The data used for model training fall within the time range of Jan 1, 2017, to Dec 31, 2019. From this data set, approximately 17 % of the cases were excluded based on the following criteria: duplicates, emergent and urgent cases, patients with age younger than 20, pregnant patients, procedure duration longer than 10 hours or less than 10 minutes and cases with missing values. The total number of cases included in the data set for model building was 142,445. This data set was then split into training (80 %) and validation (20 %) subsets for model development. Machine learning and linear regression models were developed on the training data set and validated on the validation data set using R-square and mean absolute error. The percentage of cases with duration differences falling within the threshold of minimum(maximum(15 %, 15 min),60 min) of the actual duration was also computed. Eventually, 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.

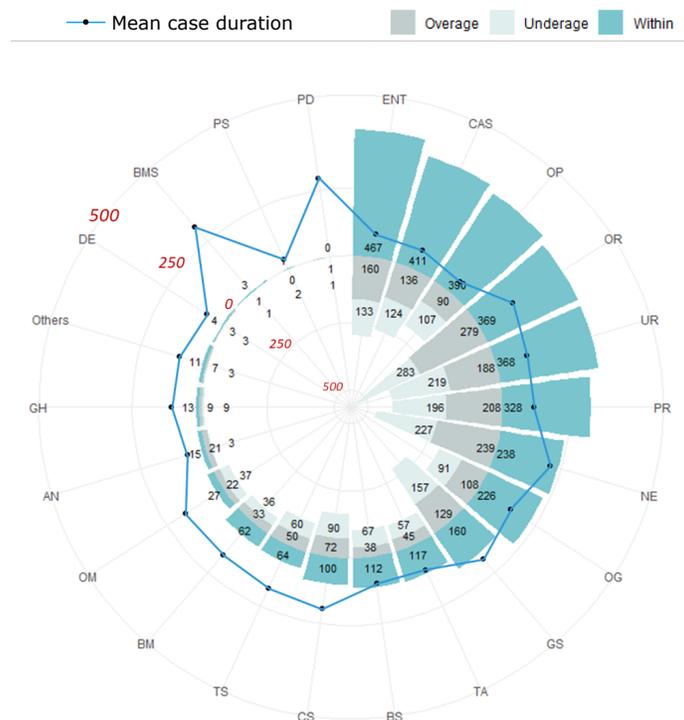


Figure S3: Numbers of cases that were predicted as overage, underage and within for each specialty in the time-wise testing set based on surgical case durations predicted by the extreme gradient boosting model. The black circle represents the mean case duration for each specialty. OP: ophthalmology; CAS: cardiovascular; ENT: otolaryngology, head and neck; TA: trauma and acute care; OG: obstetrics and gynecology; CS: colorectal; UR: urology; BM: body science and metabolic disorders; BS: breast surgical oncology; PR: plastic and reconstruction; GS: general; OR: orthopedics; TS: thoracic; NE: neurosurgery; OM: oral and maxillofacial; AN: anesthesiology; GH: gastroenterology and hepatology; DE: dermatology; PS: pediatric; BMS: bariatric and metabolic; PD: pediatric dentistry.

Model	Train set					Internal Validation set					Time-wise Test set				
	$R^2$ (%)	MAE	O (%)	U (%)	W (%)	$R^2$ (%)	MAE	O (%)	U (%)	W (%)	$R^2$ (%)	MAE	O (%)	U (%)	W (%)
Surgeon-specific	28	61.8	29	50	21	27	61.5	29	50	21	26	67.4	30	52	17
Procedure-specific	74	34.2	23	34	43	74	34.3	23	35	42	74	36.8	22	37	41
Reg	79	31	22	30	48	78	31.4	22	31	47	78	33.6	20	35	45
logReg	82	30.9	26	22	52	81	31.2	26	23	51	82	33.3	23	27	50
RF	92	18.7	17	11	73	79	28.8	25	21	54	79	31.04	24	23	53
XGB	85	27.9	25	20	55	83	28.9	25	22	53	84	31.1	23	25	51

Table S1: Performance of all the models in the training and internal validation and time-wise testing sets. The models that were included for comparison in this study were average models for surgeon- or procedure-specific scenarios, multivariate linear regression (Reg), log transformed multivariate linear regression (logReg), random forest (RF) and extreme gradient boosting (XGB). MAE: mean absolute error; U: underage; O: overage; W: within

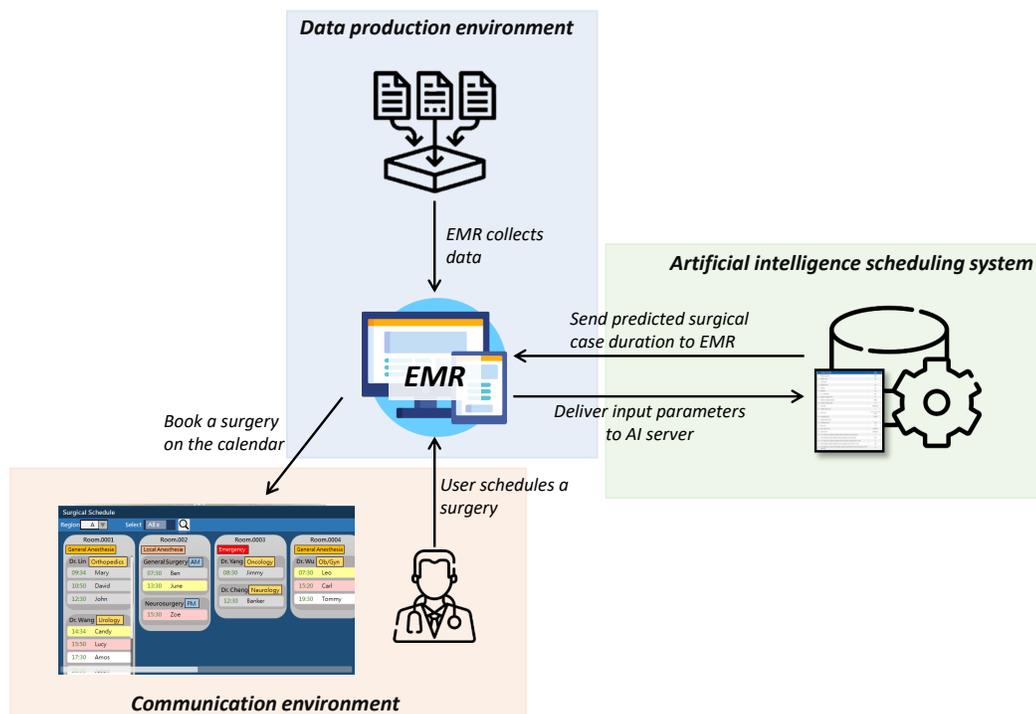


Figure S4: The mode of deployment of the machine learning (ML) model developed in this study. The ML model will be deployed as a stand-alone artificial intelligence (AI) server connected by the electrical medical record (EMR) system of the hospital. When a user schedules a surgery using the EMR system, the EMR delivers input parameters to the AI server, where a prediction of surgical case duration is made, and then the AI server sends the output back to the EMR as a suggestion for the user to book a surgery on the calendar.

Predictor variables	Correlation coefficient	Standard error	t-value	p-value
No. of previous surgeries performed by the surgeon within 24 hr	$-9.14 \times 10^{-3}$	$7.5 \times 10^{-4}$	-12.19	$<2 \times 10^{-16}$
Total surgical minutes performed by the surgeon within 24 hr	$6.48 \times 10^{-5}$	$6.22 \times 10^{-6}$	10.41	$<2 \times 10^{-16}$
No. of previous surgeries performed by the surgeon within the last 7 days	$-1.37 \times 10^{-2}$	$2.58 \times 10^{-4}$	-53.35	$<2 \times 10^{-16}$
Total surgical minutes performed by the surgeon within the last 7 days	$7.71 \times 10^{-5}$	$2.27 \times 10^{-6}$	33.89	$<2 \times 10^{-16}$

Table S2: Correlation coefficient, standard error, t-value and p-value of predictor variables extracted from primary surgeons' data. This information was obtained from the log-transformed multivariate regression (logReg) model.

	Overall (n = 142,445)	Test(n = 7,231)
<b>Gender</b>		
Male	70096 (49.2%)	3618 (50%)
Female	72349 (50.8%)	3613 (50%)
<b>Age</b>		
20-45	42396 (29.7%)	2061 (28.5%)
45-65	55134 (38.7%)	2800 (38.7%)
65-80	34493 (24.2%)	1811 (25%)
>80	10422 (7.3%)	559 (7.7%)
<b>In-/out-patient</b>		
In-patient	92780 (65.1%)	4826 (66.7%)
Out-patient	49665 (34.8%)	2405 (33.2%)
<b>Hypertension</b>		
Yes	32398 (22.7%)	1676 (23.1%)
No	62792 (44%)	3181 (44%)
Unknown	47255 (33.1%)	2374 (32.8%)
<b>BMI</b>		
<18.5	33581 (23.6%)	1945 (26.9%)
18.5-22.9	45338 (31.8%)	2470 (34.2%)
23-26.9	42392 (30%)	2246 (31.1%)
>= 27	8235 (5.8%)	439 (6.1%)
Missing	12899 (9.1%)	131 (1.8%)

Table S3: Proportion based on patient characteristics in the overall original data set and time-wise testing set.