medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

A Machine Learning Study to Improve Surgical Case Duration Prediction

Ching-Chieh Huang[1,†], Jesyin Lai[1,†], Der-Yang Cho[1], Jiaxin Yu[1,*]

1 Artificial Intelligence Innovation Center, China Medical University Hospital, Taichung, Taiwan

† C. Huang and J. Lai contributed equally to this work
*** Corresponding author: J. Yu, Artificial Intelligence Innovation Center, China Medical University Hospital, No. 2, Yuh-Der Road, 404, Taichung. Email: jiaxin.yu@mail.cmuh.org.tw**

Abstract

Predictive accuracy of surgical case duration plays a critical role in reducing cost of operation room (OR) utilization. The most common approaches used by hospitals rely on historic averages based on a specific surgeon or a specific procedure type obtained from the electronic medical record (EMR) scheduling systems. However, low predictive accuracy of EMR leads to negative impacts on patients and hospitals, such as rescheduling of surgeries and cancellation. In this study, we aim to improve prediction of operation case duration with advanced machine learning (ML) algorithms. We obtained a large data set containing 170,748 operation cases (from Jan 2017 to Dec 2019) from a hospital. The data covered a broad variety of details on patients, operations, specialties and surgical teams. Meanwhile, a more recent data with 8,672 cases (from Mar to Apr 2020) was also available to be used for external evaluation. We computed historic averages from EMR for surgeon- or procedure-specific and they were used as baseline models for comparison. Subsequently, we developed our models using linear regression, random forest and extreme gradient boosting (XGB) algorithms. All models were evaluated with R-squre (R^2), mean absolute error (MAE), and percentage overage (case duration > prediction + 10 % & 15 mins), underage (case duration < prediction - 10 % & 15 mins) and within (otherwise). The XGB model was superior to the other models by having higher R^2 (85 %) and percentage within (48 %) as well as lower MAE (30.2 mins). The total prediction errors computed for all the models showed that the XGB model had the lowest inaccurate percent (23.7 %). As a whole, this study applied ML techniques in the field of OR scheduling to reduce medical and financial burden for healthcare management. It revealed the importance of operation and surgeon factors in operation case duration prediction. This study also demonstrated the importance of performing an external evaluation to better validate performance of ML models.

Introduction

It becomes more and more important for clinics and hospitals in managing resources for critical cares during the COVID-19 pandemic. Statistics show that approximately 60 % of patients admitted to the hospital will need to be treated in the Operation Room (OR) [11], and the average cost of OR is up to 2,190 dollars per hour in the United

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

States [1,6]. 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 [6,10]. Based on these statistics, a good OR schedule and management is not only critical to patients who are in need of elective, urgent and emergent operations, but is also important for surgical teams to be prepared. Owing to the importance of OR, improvement of OR efficiency has high priority so that the cost and time spent on OR is minimized while the utilization of OR is maximized to increase surgical case number and patient access [15].

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 operations, surgical case scheduling and etc [20]. Although the process of OR is complex and involves multiple parties, one way to enhance OR efficiency is by increasing the accuracy of 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 implement negative impact on the patient, staffs and hospital [21]. In contrast, high efficiency in OR scheduling not only contribute to better arrangement for the usage of operating room and resources, it can also lead to cost reduction and revenue increment since more surgeries can be performed.

Currently, most hospitals schedule surgical case duration by employing estimations from surgeon and/or averages of historical case durations, and studies show that both of these methods have limited accuracy [14,17]. For case length estimated by surgeons, factors including patient conditions, anesthetic issues might not be taken into consideration. Moreover, underestimation of case duration often occurs as surgeon estimations were usually made by leaning towards maximizing block scheduling to account for potential cancellations and cost reduction. Furthermore, operations with higher uncertainty and unexpected findings during operation add difficulties and challenges into case length estimation [14]. Historic averages of case duration for a specific surgeon or a specific type of operation 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 large variability and lack of same combination in the preoperative data available on the case that is being performed [25].

In order to improve the predictability, researchers utilized linear statistical models, such as regression, or simulation for surgical duration prediction and evaluation of the importance of input variables [8,12,13]. However, a common shortcoming of these studies is that relatively lesser input variables or features were used in their models due to the limitation of statistical techniques in handling too many input variables. Recently, machine learning (ML) has 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 predict surgical durations [16]. Ensemble classifiers, implementing least-squares boosting and bagging models with ML, developed by Shahabikargar et al. (2017) were shown to reduce error by 55 % as compared to the original error [21]. With the use of boosted regression tree, Zhao et al. (2019) increased the percentage of accurately booked cases for robot-assisted surgery from 35 % to 52 %. Bartek et al. (2019) reported that they were able to improve predicted cases within 10 % threshold tolerance from 32 % to 39 % using an extreme gradient boost model [2]. 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 their models.

In this study, we obtained more than 170,000 cases from China Medical University Hospital (CMUH) containing 422 types of procedures across 25 different specialties. From the original data, we further analyzed the working time of primary surgeons and computed their total number of previous surgeries and the total minutes spent on

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

previous surgeries within the same day 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, especially within the same day, 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 operation case duration. In addition, number of urgent and emergent operations prior to the case that was being performed by the same surgeon, which has not been considered in other studies before, was taken into consideration as well. This factor could affect operation case duration as urgent and emergent operations happen unexpectedly and delay the start of subsequent planned surgeries. As a whole, we hypothesize that these features impose significant influences on operation case duration and may aid in improving the performance of a trained ML model.

Methods

Data sources

Data for this study were collected from EMR scheduling system of CMUH located at Taichung, Taiwan. The data set covered a broad variety and details about patients, operations, specialties and surgical teams. A total of 170,748 cases performed between Jan 1, 2017 to Dec 31, 2019 were used for model development. Meanwhile, 8,672 cases performed between Mar 1 to April 30, 2020 were used as data for external model evaluation in this study. Over 400 different types of procedures across 25 surgical specialties were included in the data set. An institutional review board approval (CMUH109-REC1-091) was obtained from CMUH before carrying out this study.

0.1 Exclusion criteria, data processing and feature selection

Emergent and urgent surgical cases were removed since these two types of operation can not be scheduled until they happen. Surgeon's age younger than 28 years and surgical case duration more than 10 hours or less than 10 minutes were also removed. Surgical records with missing values were excluded. Patients who were pregnant or underwent two or more surgical procedures at the same time or with age under 20 year-old were deleted. The exclusion criteria were shown in Fig. 1. This resulted in a data set of 142,448 cases that were used for model training and testing. The same criteria were also applied to the data of Mar 1 to April 30, 2020 and 7,231 cases remained after exclusion.

Features were selected from available data sources, based on literature review and discussion with surgeons and administrators of CMUH. Although model performance could be enhanced by some postoperative information (e.g. total blood loss), they cannot be used as features for model training because these parameters were either missing or simply estimated by surgeons before surgery. Therefore, only variables that are available before operation were selected for model development.

When visualizing all the categories of procedure types and 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, we combined categories which had case numbers less than 50 in the training set into a category and named it as 'Others'. Similarly, we combined categories for primary surgeon's ID, specialty, anesthesia type and room number which had case numbers less than 50 into the category of 'Others'.

In addition, since operation case duration can be related to the performance of surgeons and surgeons' performance is affected by their working time, we also analysed

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

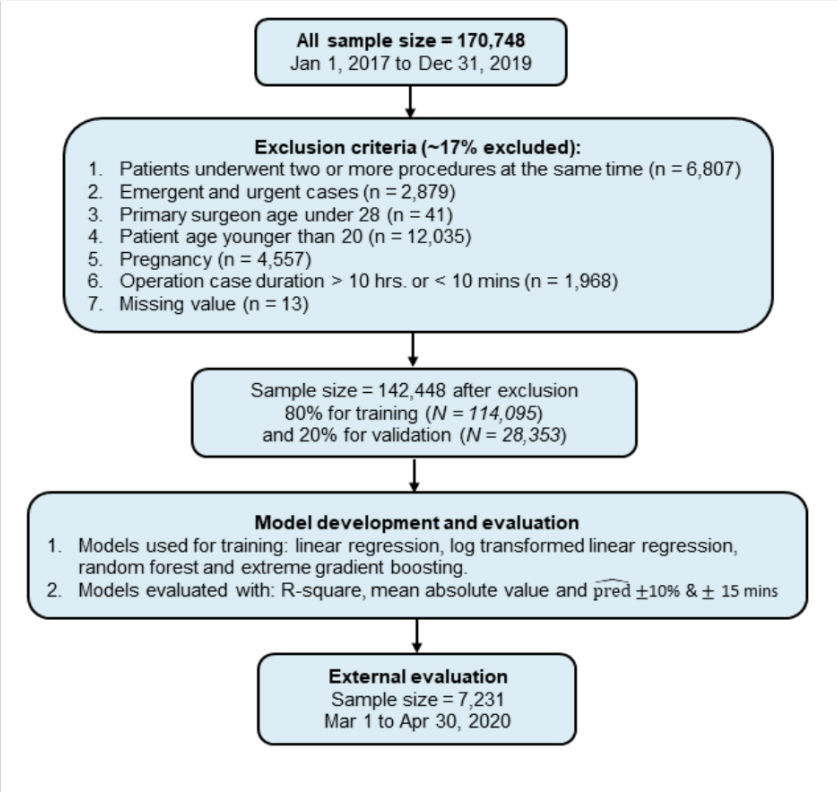

Figure 1. The workflow of model training for this study. The data set used for model training fall within the time range of Jan 1, 2017 to Dec 31, 2019. From this data set, about 17 % of the cases were excluded based on these criteria: patients with two or more surgical procedures performed at the same time, emergent and urgent cases, surgeons with age under 28, patients with age younger than 20, pregnant patients, procedure duration longer than 10 hours or less than 10 minutes and cases with missing value. The total number of cases included in the data set for model building was 142,448. 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. Percentage of cases with actual duration differences falling within 10 % and 15 minutes of predicted procedure 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Patient	Surgical team	Operation	Facility	Primary Surgeon's Prior Events
Age	Primary surgeon's ID	Procedure type	Room No.	No. of previous surgeries performed by the surgeon on the same day
Gender	Surgeon team size	Subprocedure type	Day of the week	Total surgical minutes performed by the surgeon on the same day
ICD code	Specialty	Anesthesia type	Time of day	No. of previous surgeries performed by the surgeon within the last 7 days
In- /out-patient	Primary surgeon's gender			Total surgical minutes performed by the surgeon within the last 7 days
ASA status	Primary surgeon's age			No. of previous urgent and emergent surgeries performed by the same surgeon on the same day
Hypertension				
Anemia				
Diabetes				

Table 1. Preoperative data with 24 predictor variables were used for model development. The predictor variables can be categorized by relationship to patient, surgical team, operation, facility and surgeon's prior events. ICD: International Classification of Diseases; ID: Identifier: ASA: American Society of Anesthesiologists

primary surgeons' previous surgical events. The number of previous surgeries and total surgical minutes performed by the same primary surgeons on the same day as well as within the last 7 days, and the number of urgent and emergent operations prior to the case that was being performed by the same surgeon were included in the analysis. Together, 24 predictor variables were included for predictive model building in this study. These predictors can be categorised into 5 groups: patient, surgical team, operation, facility and primary surgeon's prior events (see Table 1).

Model development and training

We applied multiple ML methods for operation case duration prediction. Operation case duration (in minutes) is the total period starting from the time patient entering into the OR to the time exiting the OR. Historic averages of case durations based on surgeon-specific or procedure-specific from EMR systems were used as baseline models for comparison in case duration prediction. At the beginning, we performed multivariate linear regression (Reg) to predict operation case duration. However, when we looked at the distribution of operation case duration, it was observed to be skewing to the right (Fig. 2). We performed logarithmic transformation on operation case duration to reduce the skewness. The model built from log transformed multivariate linear regression (logReg) outperformed Reg in all evaluation indexes. Subsequent ML algorithms were also trained by using the log transformed case duration as the target.

The first ML algorithm that we tested is random forest (RF), a tree-based supervised learning algorithm. RF uses bootstrap aggregation or 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 [19]. Bagging technique is unlikely to over-fitting, in other words, it reduces the variation without increasing the bias. Tree-based techniques were suitable for our data since they include a large number of categorical variables, e.g. ICD code and procedure type, most of which were sparse. The number of trees that was set in study is 50. Extreme Gradient Boosting (XGB) algorithm is the other supervised ML algorithm that was tested for comparison to RF. Recently, XGB algorithm gains popularity within the data science community due 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 but more computationally efficient for real-time implementation than RF. XGB and RF algorithms are different in the way of how the trees are built. It has been shown that XGB performs better than RF if parameters are tuned carefully, otherwise it would be more likely to over-fitting if the data are noisy [3, 9]. We adopted 5-fold cross validation strategy to tune out the best

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

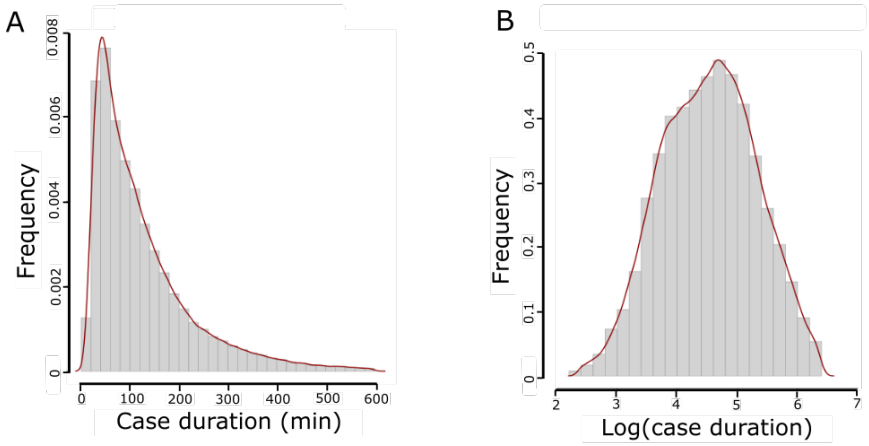

Figure 2. Log transformation of case duration converted the distribution of operation case duration from (A) skewing to the right to (B) a more normal distribution.

number of iterations by using $\eta = 0.5$ (step size shrinkage to prevent over-fitting), maximum 3 depths of the tree, $\gamma = 0.3$ (minimum loss reduction, where a larger γ represents a more conservative algorithm) and $\alpha = 1$ (L1 regularization weighting term, where a larger value indicates a more conservative model).

Data-splitting strategy was used in the training for all the models to prevent over-fitting consequences. We randomly separated the data into training and testing subsets at a ratio of 4:1. The training data were used to build different predictive models as well as to extract important predictor variables. The testing data were used for internal evaluation of the models.In addition to interval evaluation, external evaluation on all the models were performed using data from Mar 1 to Apr 30, 2020. These data were not included in the original data set for ML model training. The results obtained from external evaluation are thus better in showing the robustness of the trained model in making accurate prediction. Historic averages of case duration for surgeon- or procedure-specific calculated from EMR were also evaluated on the same internal and external testing sets to ensure fair and uniform comparison across all models. Data processing and cleaning as well as model development in this study were performed using R software. The packages "xgboost and "randomforest were used to implement XGB and RF algorithms in R [4, 5].

Model evaluation

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

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

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

Mean Absolute Error (MAE) measures the average of errors between the actual case

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

	Train set					Internal Test set					External Test set				
Model	R^2 (%)	*MAE*	*U (%)*	*O(%)*	*W (%)*	R^2 (%)	*MAE*	*U (%)*	*O(%)*	*W (%)*	R^2 (%)	*MAE*	*U (%)*	*O(%)*	*W (%)*
Surgeon-specific	31	60.4	50	31	19	30	60.3	50	31	19	30	64.6	52	32	17
Procedure-specific	68	37.3	37	25	37	66	38	38	25	36	66	40.7	40	25	35
Reg	80	30.3	32	25	43	78	31.1	32	25	42	79	33	36	23	41
logReg	84	28.7	24	27	49	83	29.5	25	27	48	84	31	27	25	47
RF	83	29.2	23	27	49	83	28.9	23	27	50	84	30.9	26	26	48
XGB	85	27.3	23	26	51	84	28.7	23	27	49	85	30.2	27	25	48

Table 2. Performance of all the models in the training, internal and external testing sets. The models that were included for comparison in this study were average models for surgeon- or procedure-specific, 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

durations and the predictions.

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \widehat{y_i}|}{N} \tag{2}$$

Percentage overage indicates the percentage of cases with actual case duration > prediction + 10 % tolerance threshold (i.e. 1.1 $*$ prediction) and prediction + 15 minutes. Meanwhile, percentage underage is the percentage of actual case duration < prediction - 10 % tolerance threshold (0.9 $*$ prediction) and prediction - 15 minutes. Therefore, percentage within equals to 100 %-(percentage overage + percentage underage).

Results

Approximately 17 % cases were excluded from the original data of Jan 1, 2017 to Dec 31, 2019 based on the exclusion criteria mentioned in Fig. 1. Therefore, 142,448 cases containing more than 420 procedural categories and 25 specialties were included for predictive model development and evaluation. Furthermore, a recent data collected from Mar 1 to April 30, 2020 (7,231 cases after exclusion) were used in external evaluation to study the robustness of model in making prediction.

Table 2 shows the results of the metrics used to evaluate performance of all the models in this study. The average model for surgeon-specific had the highest percentage underage, which is 50 % on training, internal and external testing sets. This indicates that 50 % of actual case durations were 15 minutes and 10 % lower than predicted case duration. Other metrics (R^2 and MAE) also show that the average model based on a specific surgeon was not a good estimate for operation case duration. On the other hand, the average model based on a specific procedure had lower percentage underage and overage compared to the surgeon-specific model. These differences were due to an extensive procedure classification in the procedure-specific model. However, the percentage underage was still quite high. Since no other information is taken into consideration in the average model, except durations of operation cases happened in the past, prediction bias and low accuracy usually result from the average model.

We first fitted the Reg model by including all the input variables showed in Table 1. The evaluation metrics reported lower percentage underage and higher percentage within when compared to the average model on training, internal and external testing sets (Table 2). There was a large improvement in R^2 value indicating that predictive performance of model increases when other information are taken into consideration during model development. Since the results of percentage underage, overage and within on training, internal and external testing sets were similar, over-fitting was not

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

	Actual	Average (surgeon)	Average (procedure)	Reg	logReg	RF	XGB
Total minutes	920,374	899,510	918,061	934,333	885,784	874,528	888,908
Total prediction error in minutes		467,548	294,137	238,862	224,700	223,686	218,415
Inaccurate percent (%)		50.8	32	26	24.4	24.3	23.7

Table 3. Extreme Gradient Boosting (XGB) 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.

likely to happen in the Reg model. A model is probably considered to be over-fitting when its performance is better on training set but poor on testing set.

When we log transformed operation case duration and re-ran a regression model (i.e. logReg), the performance of logReg model improved and outperformed Reg model. Since we are predicting operation case duration, log transformation of the target prevents us from getting values of zero or negative from the predicted output of the model. Log transformation has been used commonly in other studies for the same reason [12, 23]. Again in the logReg model, the results of all the evaluation metrics were close for training, internal and external training sets, so the model was not over-fitting.

Although performance of the logReg model was not bad, an assumption of linear relationship between target and input variables was applied in both the Reg and the logReg models. The relationship between target and input variables is usually non-linear in a real world situation. ML algorithms are helpful in making prediction in a more complicated scenario. RF model is the first ML model we built in this study. There was a slight improvement in the performance of RF model when comparing the results of all evaluation metrics to those of the logReg model on both internal and external testing sets. An XGB model was developed subsequently because training duration of the RF model was time consuming and caused low computing efficiency. Performance of the XGB model was better than the RF model on training set but did not improve a lot compared to the RF model on internal and external testing sets. Since XGB was more computing efficient than RF, the XGB model was chosen to be the best model and was used in subsequent analysis.

In addition to the three key metrics, we studied inaccuracy of different models by using external testing set. We calculated the total prediction error (in minutes) and the corresponding inaccurate percentage for all the models. The results are reported in Table 3. Total minutes of actual represent the sum of operation case durations for 7,231 cases in the external testing set. Inaccurate percent was derived from the percentage of total prediction error divided by actual total minutes. The outcome shows that inaccurate percent of the XGB model was the lowest among all the models. Inaccurate percent of the XGB model was also more than 50 % lower than that of the average model for surgeon-specific and about 25 % lower than that of the procedure-specific average model. This implies that prediction made by the XGB model had low inaccuracy and might help to increase the efficiency of OR scheduling.

In Fig. 3, we plotted scatter plots of actual versus predicted duration on the external testing set for the average models of surgeon- and procedure-specific, and the XGB model. A straight line indicating the theoretical perfect relationship, i.e. predicted and actual procedure duration are identical, was added as a reference in each scatter plot. The data points of the XGB models were aligned closer to the straight line. Therefore, the XGB model showed a higher correlation between predicted and actual duration compared to the other two types of average model. Fig. 4 shows the density plot of differences between actual and predicted case durations for the two average models and the XGB models. It clearly demonstrates that the error distribution of XGB model was narrower and closer to 0. As a result, the XGB model is more accurate than the other models in predicting operation case duration.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

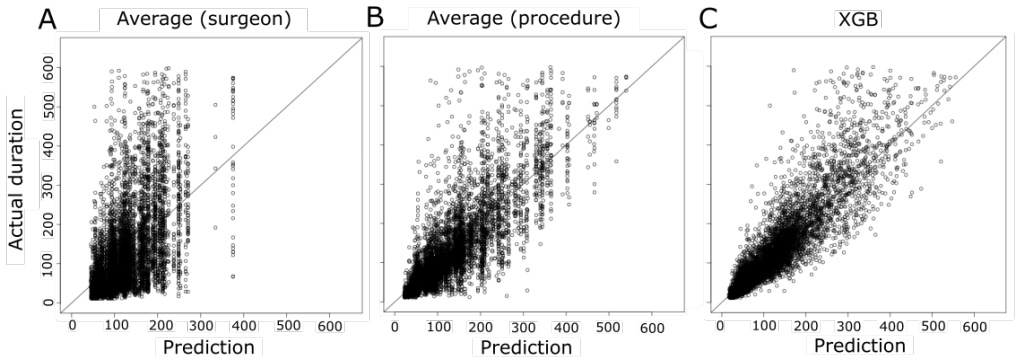

Figure 3. Scatter plots of actual duration versus prediction obtained from average models for (A) surgeon- and (B) procedure-specific, as well as from (C) Extreme Gradient Boosting (XGB) model. A straight line with correlation = 1, representing a perfect relationship between predicted and actual values, was added as a reference in each plot.

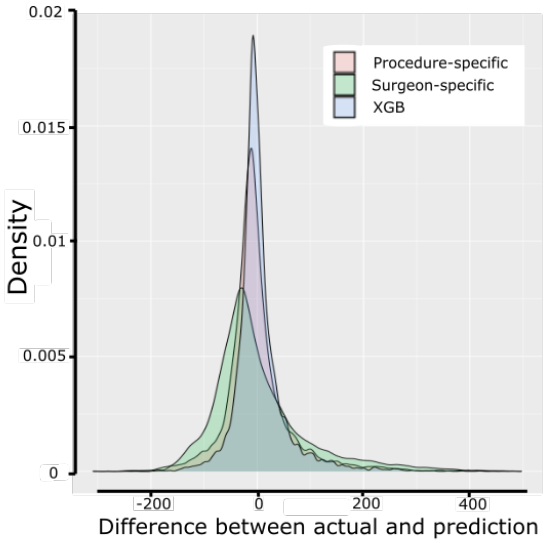

Figure 4. Density plot of differences between the actual operation case durations and predicted case durations obtained from the XGB model (light blue color) was narrower and centered more at 0 than density plots of those obtained from the average models (pink and cyan colors). In the average models, previous operation case durations, either averaging for a specific surgeon (cyan color) or specific procedure (pink color), were used as predictions.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Variable	Weighted feature gain (%)	Data type	No. of categories
Anesthesia type	53.11	Categorical	8
Procedure type	12.16	Categorical	422
Hypertension	11.62	Categorical	3
Subprocedure type	6.86	Categorical	394
Primary surgeon's ID	3.47	Categorical	184
In- /out-patient	3.41	Categorical	2
Specialty	3.13	Categorical	25
Room No.	1.50	Categorical	53
ICD code	0.95	Categorical	1229
No. of previous surgeries performed by the surgeon within the last 7 days	0.80	Numerical	-
Total surgical minutes performed by the surgeon within the last 7 days	0.79	Numerical	-
Primary surgeon's age	0.64	Numerical	-
Surgeon team size	0.57	Categorical	5
Total surgical minutes performed by the surgeon on the same day	0.31	Numerical	-
ASA status	0.23	Categorical	7

Table 4. Top 15 important features used by the Extreme Gradient Boosting (XGB) model to predict operation case durations are shown in the table. Total surgical minutes and number of previous surgeries performed by the surgeon within the last 7 days or on the same day were included in this top 15 list. Weighted feature gain in % is an output of the XGB algorithm. ICD: International Classification of Diseases; ID: Identifier: ASA: American Society of Anesthesiologists

To visualize variable importance in the XGB model, we extracted the weighted feature gain (WFG) from the model. WFG was computed based on the reduction of model accuracy when the variable was removed. This value serves as an indication on how important the variable is in making a branch of a decision tree to be purer [5, 22]. A higher WFG percentage indicates that the variable is more important. The result of the top 15 important variables are shown in Table 4. One thing worth noting is that 3 of the top 4 important variables are attributed to operation information. Moreover, three of the features which we computed from surgeons' data (i.e. total surgical minutes performed by the surgeon within the last 7 days and on the same day, and number of previous surgeries performed by the surgeon within the last 7 days) were included in this top 15 list.

Discussion

Accurate prediction of operation case duration is vital in elevating OR efficiency and reducing cost. This study not only helps to improve accuracy of OR case prediction, it also has novelty in the following aspects. First, the data set used in this study contained more than 140,000 cases and more than 400 different types of surgical procedures which set up a new benchmark for huge amount and large diversity. The maximal number of cases that had been used in other studies were in the range of 40,000 to 60,000 [2, 21]. Second, OR events was modeled as dependent events instead of independent. To this end, we extracted some additional information from surgeons' data, e.g. previous working time and number of previous surgeries of the primary surgeons within the last 7 days and on the same day, and these information were taken into consideration during model building. Third, we tested the model on daily clinical workflows from Mar to

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

April 2020 as external testing data for model evaluation. Fourth, though urgent and emergent surgeries were excluded from the data, number of urgent and emergent operations prior to the case that was being performed by the same surgeon was included as an input variable to account for its effect on operation case duration.

Currently, surgical cases at CMUH are scheduled according to estimates made by primary surgeons. However, surgeon estimates rely heavily on prior experiences of the surgeons and many factors beyond expectation will not be taken into consideration. Since there is no formal record on surgeon estimates, we used averages calculated based on a specific surgeon or procedure type on the testing set to be our baseline models. The performance of these two average models, as reported in Table 2, clearly showed that these models were poor in predicting operation case duration. They also tended to under-predict operation case duration according to their scatter plots of actual versus prediction and density plot of differences between actual versus prediction (see Fig. 3 and 4). When 24 feature variables (Table 1) were included in our model development, R^2, MAE, percentage underage, overage and within improved greatly compared to the baseline models. We applied 15 minutes as tolerance threshold for percentage underage, overage and within because $\pm$ 15 minutes is an acceptable periodic range in CMUH to be considered as accurately booking. To avoid having too stringent standard and to better compare our outcomes with other studies [2, 24], tolerance threshold of 10 % was also applied.

By using regression and ML approaches, we were able to decrease the total prediction error (Table 3) of operation case durations at CMUH. Among all the models, performance of the XGB model was considered to be the best because it was more computing efficient and had the lowest inaccuracy. Moreover, even though the results of evaluation metrics of the RF model were similar to the XGB model, the XGB model was still able to reduce the total prediction error in minutes from 223,686 to 218,415 minutes. In other words, the XGB model was able to save more than 5,000 minutes of idle or delay times than the RF model. Since most ORs usually have multiple cases scheduled per day, the total prediction error represents the cumulative effect of total OR cases in the 2-month period of Mar to April 2020. This cumulative effect may eventually reflects a significant financial advantage in scheduling an additional operation case [7]. This would also lead to a significant cost reduction and increment in revenue because ORs are utilized appropriately and efficiently.

It has been reported in the past studies that primary surgeons contributed the largest variability in operation case duration prediction compared to other factors attributed to patients [2, 16, 23]. These studies provide evidence and rationale that more factors relating to primary surgeon should be added as input variables in the training of ML models. Moreover, extensive feature engineering usually improve the quality of ML model which can be independent to the modeling technique itself. As a result, in addition to primary surgeon's identifier, gender and age, we computed previous working time and number of previous surgeries performed by the same primary surgeons within the last 7 days and on the same day. We also counted the number of urgent and emergent operations prior to the case that was being performed by the same primary surgeon. These variables extracted from the data of primary surgeon were significantly ($p < 0.05$) correlated with operation case duration (see Table 5 in Appendix). The correlation coefficients of these variables also revealed that an operation case duration performed by a primary surgeon may decrease as he or she becomes more familiar with the surgical procedure but may increase if his or her total surgical minutes are too long. Although performing a surgery multiple times on different patients may help a primary surgeon to be more efficient in his or her next operation, long working time may also lead to lethargic and affect the primary surgeon's performance.

In the methodology of data processing, for predictor variables which contained a lot

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

of categories, we grouped categories that had cases less than 50 into a categories named 'Others'. In addition to reducing data dimensionality for categorical features, this may aid in generalization of our model. This indicates that our model will still be able to predict case duration even for operations that are rare. Moreover, our model can be applied to new primary surgeons, who are not included in the training set during model development, by setting their ID as 'Others' for case duration prediction. However, there is still a need to update our model after a while, for example, when the operation cases performed by a new primary surgeon has increased beyond a certain number. In terms of timing, we recommend updating the model annually by using operation cases performed in the most recent 3 years as training data.

One limitation in this study is that we selected predictor variables which could only be extracted from preoperative data. Our ML model still needs to be improved in order to be able to predict surgical case duration dynamically. For example, blood loss during operation may affect case duration as an unexpected increase in blood loss may cause surgeons to take longer time to complete the surgery. Therefore, it would be better if intra-operative data are incorporated during ML model development and prediction made by the ML model can be updated during operation. One common issue in all ML studies in predicting operation case duration, including our study, is that ML models were developed using data from a single site. These ML models have difficulties in generalization, since the surgical team, facilities and patient populations are different across entities. It has to be custom made for a given organization using training data containing its patients, procedures, surgeons, medical staffs, and the facility itself. As a result, the exact same ML model is not meant to and will not perform well when applied to another organization or hospital. The other interesting issue of applying ML or artificial intelligence in operation estimation is that medical technologies evolve fast. Hence, how frequent should a ML or artificial intelligence model need to be updated still remains to be answered.

Conclusion

The XGB model was superior in predictive performance when comparing to the average, the Reg and the logReg models. The total inaccuracy of predicted outcomes of the XGB model was the lowest among the other models developed in this study. Although the performance of the RF model was close to the XGB model, the XGB model was more computing efficiency than the RF model in which it took shorter time to complete the training process. The coefficient of determination (R^2) was higher while percentages of under- and over-prediction of the XGB model built in this study were also lower than other ML studies [2, 21, 24]. Moreover, this model improves the current OR scheduling method which is based on estimates made by surgeons at CMUH.

We propose extracting additional information from operation and surgeons' data to be used as predictor variables for ML algorithm training since their importance was high in the XGB model. Moreover, we validated the model types using an external testing set in additional to the internal testing set split from the original data used in model training. This helped us to validate and test the models in a more stringent and rigorous way. Therefore, we suggest external evaluation should be used as a tool to better validate the predictive power of ML models in the future.

Acknowledgments

The authors would like to thank Shu-Cheng Liu, Jhao-Yu Huang and Min-Hsuan Lu in providing feedback during the progress of this study.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. S. Barbagallo, L. Corradi, J. De Ville De Goyet, M. Iannucci, I. Porro, N. Rosso, E. Tanfani, and A. Testi. Optimization and planning of operating theatre activities: An original definition of pathways and process modeling. *BMC Medical Informatics and Decision Making*, 15(1), 5 2015.

2. M. A. Bartek, R. C. Saxena, S. Solomon, C. T. Fong, L. D. Behara, R. Venigandla, K. Velagapudi, J. D. Lang, and B. G. Nair. Improving Operating Room Efficiency: Machine Learning Approach to Predict Case-Time Duration. In *Journal of the American College of Surgeons*, 2019.

3. C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz. A Comparative Analysis of XGBoost. 11 2019.

4. L. Breiman, A. Cutler, A. Liaw, and M. Wiener. Package 'randomForest' Title Breiman and Cutler's Random Forests for Classification and Regression. 2018.

5. T. Chen, T. He, M. Benesty, and V. Khotilovich. Package 'xgboost' Type Package Title Extreme Gradient Boosting. 2020.

6. C. P. Childers and M. Maggard-Gibbons. Understanding costs of care in the operating room. *JAMA Surgery*, 153(4), 4 2018.

7. F. Dexter and A. Macario. Decrease in Case Duration Required to Complete an Additional Case During Regularly Scheduled Hours in an Operating Room Suite. *Anesthesia & Analgesia*, 88(1):72–76, 1 1999.

8. M. J. Eijkemans, M. Van Houdenhoven, T. Nguyen, E. Boersma, E. W. Steyerberg, and G. Kazemier. Predicting the unpredictable: A new prediction model for operating room times using individual characteristics and the surgeon's estimate. *Anesthesiology*, 112(1):41–49, 2010.

9. J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 2001.

10. B. M. Gillespie, W. Chaboyer, and N. Fairweather. Factors that influence the expected length of operation: Results of a prospective study. *BMJ Quality and Safety*, 21(1):3–12, 1 2012.

11. T. Gordon, S. Paul, A. Lyles, and J. Fountain. Surgical unit time utilization review: Resource utilization and management implications. *Journal of Medical Systems*, 12(3):169–179, 6 1988.

12. N. Hosseini, M. Y. Sir, C. J. Jankowski, and K. S. Pasupathy. Surgical Duration Estimation via Data Mining and Predictive Modeling: A Case Study. Technical report, Mayo Clinic.

13. P. Kougias, V. Tiwari, and D. H. Berger. Use of simulation to assess a statistically driven surgical scheduling system. *Journal of Surgical Research*, 201(2):306–312, 4 2016.

14. D. M. Laskin, A. O. Abubaker, and R. A. Strauss. Accuracy of predicting the duration of a surgical operation. *Journal of Oral and Maxillofacial Surgery*, 71(2):446–447, 2 2013.

15. W. C. Levine and P. F. Dunn. Optimizing Operating Room Scheduling, 2015.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

16. N. Master, Z. Zhou, D. Miller, D. Scheinker, N. Bambos, and P. Glynn. Improving predictions of pediatric surgical durations with supervised learning. *International Journal of Data Science and Analytics*, 4(1):35–52, 8 2017.

17. J. H. May, W. E. Spangler, D. P. Strum, and L. G. Vargas. The Surgical Scheduling Problem: Current Research and Future Opportunities. *Production and Operations Management*, 20(3):392–405, 5 2011.

18. D. Nielsen. Tree Boosting With XGBoost - Why Does XGBoost Win "Every" Machine Learning Competition? *Tree Boosting With XGBoost - Why Does XGBoost Win "Every" Machine Learning Competition?*, 2016.

19. A. M. Prasad, L. R. Iverson, and A. Liaw. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, 9(2):181–199, 3 2006.

20. D. H. Rothstein and M. V. Raval. Operating room efficiency. *Seminars in Pediatric Surgery*, 27(2):79–85, 4 2018.

21. Z. Shahabikargar, S. Khanna, A. Sattar, and J. Lind. Improved Prediction of Procedure Duration for Elective Surgery. *Studies in health technology and informatics*, 239:133–138, 2017.

22. Y. Y. Song and Y. Lu. Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2):130–135, 4 2015.

23. D. P. Strum, A. R. Sampson, J. H. May, and L. G. Vargas. Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology*, 92(5):1454–1466, 2000.

24. B. Zhao, R. S. Waterman, R. D. Urman, and R. A. Gabriel. A Machine Learning Approach to Predicting Case Duration for Robot-Assisted Surgery. *Journal of medical systems*, 43(2):32, 1 2019.

25. J. Zhou, F. Dexter, A. MacArio, and D. A. Lubarsky. 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*, 11(7):601–605, 11 1999.

1 Appendix

medRxiv preprint doi: https://doi.org/10.1101/2020.06.10.20127910; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Predictor variables	**Correlation coefficient**	**Standard error**	**t-value**	**p-value**
No. of previous urgent and emergent surgeries performed by the same surgeon on the same day	1.49×10^{-2}	6.71×10^{-3}	2.223	0.03
No. of previous surgeries performed by the surgeon on the same day	-8.09×10^{-3}	7.34×10^{-4}	-11.029	$<2 \times 10^{-16}$
Total surgical minutes performed by the surgeon on the same day	6.06×10^{-5}	6.10×10^{-6}	9.946	$<2 \times 10^{-16}$
No. of previous surgeries performed by the surgeon within the last 7 days	-1.82×10^{-3}	3.59×10^{-4}	-5.08	3.78×10^{-7}
Total surgical minutes performed by the surgeon within the last 7 days	1.65×10^{-5}	2.84×10^{-6}	5.81	6.22×10^{-9}

Table 5. Correlation coefficient, standard error, t-value and p-value of predictor variables extracted from primary surgeons' data. These information were obtained from the log transformed multivariate regression (logReg) model.