Published in final edited form as:

J Am Coll Surg. 2019 October ; 229(4): 346–354.e3. doi:10.1016/j.jamcollsurg.2019.05.029.

Improving Operating Room Efficiency: A machine learning approach to predict case-time duration

Matthew A. Bartek, MD MPH^{*,a}, Rajeev C. Saxena, MD MBA^{*,b}, Stuart Solomon^b, Christine T. Fong, MS^b, Lakshmana D. Behara, MS^c, Ravitheja Venigandla, MS^c, Kalyani Velagapudi, PhD^c, Bala G. Nair, PhD^b, John D. Lang, MD^b

^aDepartment of General Surgery, University of Washington, Seattle, WA

^bDepartment of Anesthesiology & Pain Medicine, University of Washington, Seattle, WA

^cPerimatics LLC, Bellevue, WA

Abstract

Background—Accurately estimating operative case-time duration is critical for optimizing operating room utilization. Current estimates are inaccurate and prior models include data not available at the time of scheduling. Our objective was to develop statistical models in a large retrospective dataset to improve estimation of case-time duration relative to current standards.

Study Design—We developed models to predict case-time duration using linear regression and supervised machine learning (ML). For each of these models, we generated: 1) service-specific models and 2) surgeon-specific models in which surgeons were modeled individually. Our dataset included 46,986 scheduled surgeries performed at our center from January 2014 to December 2017, with 80% used for training and 20% for model testing/validation. Predictions derived from each model were compared to our institutional standard. Models were evaluated based on accuracy, overage (case duration > predicted + 10%), underage (case duration < predicted – 10%), and the predictive capability of being within a 10% tolerance threshold.

Results—The ML algorithm resulted in the highest predictive capability. The surgeon-specific model was superior to the service-specific model, with higher accuracies, lower percentage of overage and underage, and higher percentage of cases within the 10% threshold. The ability to predict cases within 10% improved from 32% using our institutional standard to 39% with the ML surgeon-specific model. The majority of the information utilized in the models was based on procedure and personnel data rather than patient health status.

Corresponding Author: Matthew Bartek, MD MPH, 3424 Meridian Ave N., Apt 10, Seattle WA, 98103, Telephone: 617-794-0572 | Fax: 617-890-6885, bartek@uw.edu.

Author Contributions (See separate form as well):

Conception and Design: Bartek, Saxena, Solomon, Fong, Bair, Lang

Acquisition of Data: Fong, Bair, Lang

Model development: Behara, Venigandla, Velagapudi, Nair

Analysis and Data Interpretation: All

Manuscript Drafting: Bartek, Saxena, Nair

Critical Revision: All

*Co-first authors

Conclusion—Our study is a notable advancement towards statistical modeling of case-time duration across all surgical departments in a large tertiary medical center. Machine learning approaches may improve case duration estimations, enabling improved OR scheduling, efficiency, and reduced costs.

Precis

Accurately estimating operative case-time duration is critical for optimizing operating room utilization. Current estimates are inaccurate and prior models include data not available at the time of scheduling. Machine learning approaches may improve case duration estimations, enabling improved OR scheduling, efficiency, and reduced costs.

Keywords

Algorithms; Elective Surgical Procedures/economics; Operating Rooms/economics; Operating Rooms/organization & administration; Efficiency; Organizational; Surgery case duration

Introduction

The operating room (OR) is among the highest revenue generators and accounts for as much as 42% of a hospital's revenue.(1,2) On the other hand, it is also accounts for a high cost of use, estimated at \$36 per minute.(1, 2) Therefore, optimizing OR utilization is vital for delivering efficient and cost-effective care. A key first step towards scheduling surgical procedures is the estimation of their duration. Accurate estimation enables optimal case scheduling, appropriate allocation of resources (equipment, personnel, and facilities), and creation of efficient patient flows.(3) Inaccurate case-time estimation can result in overage – when cases last longer than anticipated beyond a set tolerance threshold, or underage – when cases last shorter than anticipated beyond that same tolerance threshold. Moreover, inefficient ORs and delays reduce staff morale and patient satisfaction.(4)

Surgical scheduling has long relied on projected case-time durations submitted by surgeons themselves. However, multiple studies have demonstrated the limited accuracy of surgeon estimates.(5–7) Incentives can drive underestimates in case duration to maximize block scheduling, at the potential cost of staff overtime and potential cancellations. Certain operations such as those for oncologic resection have higher uncertainty, and thus intra-operative findings may strongly influence case duration. In addition, there are multiple patient, anesthetic, and system factors that may not be considered in the surgeon estimation. (8, 9) Alternatively, in many electronic medical record (EMR) scheduling systems, historical averages of case-time durations for a specific surgeon have also been used, though these too have been shown to lack the required accuracy due to variations in the pre-operative data available on the case which is being performed.(10–12)

Case-time estimation at our institution is currently based on two parameters, surgeon and primary procedure. The EMR scheduling system takes the primary procedure to be scheduled and generates the average of the case-time for the previous 10 procedures performed by that surgeon. However, the surgery scheduler often overrides this calculation, and replaces it with the primary surgeon's estimation of the case duration. From interviews

with surgeons and schedulers, this practice occurs because the system calculation is perceived as overly simplistic and does not account for other variables. Namely, the primary surgeon has more experience and knowledge of the specific patient, operation, and case complexity. As a result, the surgeon's estimation of case-time duration represents the current standard.

Prior studies have sought to improve upon estimated case-time duration, though no single approach has gained wide acceptance.(4, 10, 13–20). Statistical regression models have been used to predict case-time duration and assess the relative importance of input variables.(10, 21) For example, Edelman *et al.* were able to reduce the variation in total procedure time predictions in elective ophthalmology cases by 25% by employing a linear regression model that used surgeons' pre-surgical estimates as well as the type of operation, type of anesthesia, ASA class, and patient age. Master *et al.* compared multiple machine learning (ML) techniques including decision tree regression, random forest regression, and gradient boosted regression trees as well as hybrid combinations to predict case durations. (10) However, the models were trained on only 10 operations within a single specialty, thus limiting their generalizability (22)

The majority of studies aimed at improving surgical case-time estimates have focused on a single subspecialty, which provides limited utility for a clinical administrator managing the entire set of operating room suites. Moreover, many of the models did not restrict the model inputs to only pre-operatively available information, potentially leading to lower accuracy in a prospective implementation.

We sought to improve the accuracy of case-time duration estimates using data available preoperatively. Recently, data science methods such as machine learning (ML) models have gained increasing attention for their ability to predict perioperative events and operational factors.(23) Estimating case duration is particularly well-suited for ML models given that the datasets are large, well-annotated, and potentially captures the numerous factors that may influence case duration. We developed linear regression and ML models to predict OR case-time duration. The predictions from these models were compared retrospectively to the actual case times and predictions based on surgeon schedulers. We hypothesized that both linear regression and ML models would provide more accurate case-time duration predictions than our current standard.

Methods

Definitions & Model Context

Case-time duration was defined as the total minutes from patient entry into the operating room to room exit. This duration was selected due to its importance in operating room scheduling and utilization. Accuracy for each model was defined as $1 - \text{MAPE}$ (mean absolute percentage error). "Overage" cases were those where the actual case-time duration exceeded the predicted time by greater than a 10% tolerance threshold. Similarly, cases were categorized as "underage" when the actual case-time duration was less than the prediction with a 10% tolerance threshold. "Within" cases were categorized when the actual case-time duration fell within prediction $\pm 10\%$ tolerance threshold. For short duration procedures

where predicted time was less than 100 min and the 10% tolerance was therefore less than 10 minutes, the threshold was used as 10 minutes. These categorizations provided a practical method to evaluate the predictive capabilities of a model.

The University of Washington Medical Center (UWMC) is a major tertiary care center with greater than 30 operating rooms and 500 patient beds. The hospital serves the Greater Seattle area, and has a wide catchment region including 5 states. The case complexity and average case duration are high. During the year 2017, the mean case-time duration at UWMC was 3 hours 13 minutes in 14,345 cases. Only 31% of cases were within cases, i.e. they were predicted accurately by surgeon schedulers with a 10% tolerance threshold. 42% of cases were overage cases (predictions underestimated the duration), while 27% of cases were underage cases (predictions overestimated the duration).

Data Sources

After obtaining institutional review board approval (Study 00005331), we used a perioperative electronic medical record database to obtain operating room metrics as well as patient information to develop our predictive models. Notably, several personnel and procedure variables are not available preoperatively. For example, these include anesthesiologist, anesthetic plan, nursing staff, billing CPT codes, and intra-operative events. While model performance may be enhanced with these data, they would not be relevant in a prospective implementation. Therefore, we only utilized variables that were available preoperatively in model development. Secure and de-identified preoperative data for model development were provided by the Center for Perioperative & Pain Initiatives in Quality Safety Outcome (PPiQSO) at the University of Washington.

The model building approach and inclusion/exclusion criteria are shown in Figure 1. The starting dataset was comprised of preoperative data for 4 years from January 2014 to December 2017 that included scheduled surgeries performed on weekdays for adult patients (age ≥ 18 years). Procedures performed in off-site locations such as the endoscopy suite and radiology or cardiology procedure rooms were not included. Surgeries with key missing data were excluded as shown in Figure 1. This resulted in a dataset of 46,986 procedures that was used for model development and validation. The data was randomly divided into a training data set (37,588 cases; 80% of data) and a left-out testing data set (9,398 cases; 20% of data). The machine learning and linear regression models were developed on the 80% training data and tested with the 20% dataset. The EMR and surgeon scheduler estimates were also evaluated on the same testing dataset to ensure uniform comparison across all models. Preoperative patient, procedure and personnel data parameters used in the models are categorized in Table 1. A full list of predictor variable inputs derived from the preoperative data is outlined in the Appendix Table A1.

Model Development

The model development followed a standard data science approach. (23) Categorical variables were converted into a binary representation for each category (dummies). An exploratory data analysis (EDA) was first performed to characterize the data and to inform development of predictive models. Using business knowledge and EDA results, we

identified potential model features. Exploratory univariate analysis models (linear regression) were created to identify which direct and derived features (such as historical averages of case time duration) are associated with procedure duration.

Two types of models were developed to predict case time duration, linear regression and supervised machine learning models. Model development was performed in R: A Language and Environment for Statistical Computing (Vienna, Austria). The machine learning models developed were non-parametric ensemble models that combine multiple predictive techniques to produce the strongest predictive power without overfitting the data. Two different non-parametric models were developed – random forest and extreme gradient boosting (XGBoost). Broadly speaking, random forest models are easier to tune and are more robust to overfitting while XGBoost provides faster computational speed.

We utilized two approaches to develop the prognostic models. First, we generated service-specific models with patient, surgeon, and procedure information as data inputs for each surgical specialty (e.g.: plastic surgery, otolaryngology, vascular surgery). Second, we generated a series of surgeon-specific models in which surgeons were modelled individually. Only surgeons who performed a certain minimum number of procedures (100 procedures) in the training dataset were selected for surgeon-specific models. A total of 12 service-specific models corresponding to each surgical specialty and 92 surgeon-specific models for surgeons that met the minimum procedure count threshold were developed.

Model parameters were determined based on the training data set. For linear regression models, the selection of predictor variables for model development was based on those variables that passed multicollinearity test (Variance Inflation Factor < 2.0). For the non-parametric models, all features were used because overfitting was not a major concern.

Model Evaluation

Models were used to predict case-time durations for procedures in the left-out dataset (20% of original) and the predictions were compared against the actual case-time duration. The 92 surgeon-specific models were analyzed as a single composite surgeon-specific model, as were the 12 service-line models as a composite service-specific model. Additionally, we also compared the model predictions against the scheduler and EMR estimations. The key metrics to measure model performance included:

- a. Prediction accuracy: $1 - \text{MAPE}$, where MAPE is the mean absolute percent error.
- b. Percentage Overage: Percentage number of procedures that had overage
- c. Percentage Underage: Percentage number of procedures that had underage
- d. Percentage Within: Percentage number of cases with actual case-time duration within 10% of prediction (desired target metric)

The type of model (linear, random forest, or XGBoost) and the type of approach (service-specific or surgeon-specific) that yielded the highest accuracy was selected for further evaluation. Specifically, the range and distribution of model accuracies as well as the distribution of prediction error were determined to gain further insights into model performance.

Results

The performance comparison of the proposed models and existing models (surgeon scheduler and EMR estimation) are summarized in Table 2. Among the 3 types of models, the XGBoost models had the most predictive capability, with the linear model having the least. The surgeon-specific model (composite of 92 individual models) performed better than the service-specific one (composite of 12 service-specific models) with higher accuracies, lower percentage overage and percentage underage, and higher percentage within values. Figure 2 shows histograms of the distribution in prediction error for the surgeon scheduler, EMR estimation, service-specific model, and surgeon-specific model, further illustrating the highest performance of the surgeon specific models.

Figure 3 shows the distribution of model accuracies for the 92 XGBoost surgeon-specific models. The most surgeon accurate models could predict 50% of cases accurately with a 10% tolerance threshold. The least accurate models were no worse than the surgeon schedulers. Despite surgeons overriding the EMR computerized estimates in 66% of cases – with the majority reducing the estimate case duration – prediction accuracy within 10% was only marginally better (32% vs 30%). Both estimation techniques had less predictive power relative to the XGBoost surgeon-specific results.

Predictor variables were weighted based on their percentage frequency in surgeon-specific models multiplied by the information gain when including that variable into the model. The list of features is listed in Table 3. The majority of the information utilized in the models was based on procedure and personnel data. The four variables with highest gain included: the average case-time over the prior 10 instances for a given procedure as well as a given sub-procedure, and for a given procedure performed by a given surgeon as well as a given sub-procedure performed by a given surgeon. Whether or not a patient was scheduled as an outpatient or inpatient was the fifth most important feature. Overall, patient health metrics had a much smaller role compared to personnel or procedure factors in predicting case duration.

Discussion

Accurate estimation of surgical case-time duration is critical to effective block utilization, staffing, and cost reduction. We used multiple modeling approaches to compare case-time duration predictions across surgical departments, and improve upon the current standard of surgeon scheduler estimation. The study is novel for its scope (using a large clinical dataset spanning 4 years and >47,000 cases), practical focus (limiting the data inputs for our models to only that which is available pre-operatively), and approach of developing both service-specific and surgeon-specific models.

The surgeon-specific ML models provided superior predictions when compared to service-specific ML models. In our development of service-specific models, the primary surgeon was the largest contributor to variability in the model. This gave the impetus for developing surgeon-specific models to improve prediction accuracy. This finding builds on prior work in the literature. Master *et al.* improved predictions compared to surgeons' predictions and

historical averages, and found that of all input variables, the primary surgeon was the most impactful to decreasing variation in the model. Similarly, Strum *et al.* showed that compared to patient factors and intraoperative variables such as the anesthesiology team, type of anesthesia, and procedure code, surgeons are the most important source of variability in case-time duration predictions.(16)

Among the two ML models, the XGBoost model yielded better predictions of case duration than Random Forest model. Though both models are decision tree based, the XGBoost model is more computationally efficient and thus better suited for wider, more real-time implementation. Non-ML, linear regression model performed poorly and was inferior to the surgeon scheduler and EMR estimation. We suspect this is because case estimation is not a linear problem and the assumptions concerning data characteristics for linear model development may not be valid.

Surgeon schedulers tended to systematically underestimate the case duration as seen in Figure 2. This underestimation results in overage and is particularly problematic for long cases. On the other hand, the surgeon specific models produced tighter estimates closer to the desired $\pm 10\%$ range with fewer overage and underage cases as compared to the schedulers. However, as Figure 3 highlights not all surgeon-specific models are equally accurate. For practical implementation, we can selectively choose the models with the desired accuracy to optimize overall case duration estimations.

Despite surgeons revising the EMR estimate over 2/3 of the time, their ability to predict case duration within 10% was essentially the same, 32% for surgeon scheduler versus 30% for EMR respectively. This suggests that despite the additional knowledge that surgeons have concerning their cases, it is not easy to heuristically translate this information to more accurate predictions. When evaluating the most important features in the XGBoost surgeon-specific models, the top 4 features as shown in Table 3 are variations of the average case time durations of the primary surgeon, the primary procedure, and the first subprocedure. This helps explain why the EMR estimation technique of averaging surgeon-specific case durations has worked reasonably well. The majority of information utilized for modeling derives from this fundamental case information.

We used preoperative data to estimate surgical times. However, other investigators have attempted to create real-time models of surgical case-time duration. (18) Thus, in cases where an unexpected bleed is encountered, it may be catalogued by the surgical staff and an “updated” time estimate could be generated. There may be opportunities to build on our current models by incorporating real-time changes into a model.

There are multiple limitations in our study. Our ML models were developed at one institution which is a referral center with high acuity patients; important determinants of case-time duration may differ at community centers and other hospital settings. While the models can improve estimation of case duration, this does not necessarily mean that OR utilization on the whole will improve. Both the clustering and the amount of time gained by improved estimation affects the ability to add revenue generating activities such as scheduling additional cases, or reducing costs such as overtime staffing. Modeling the

economic ramifications of our ML models is a nuanced endeavor with multiple considerations. Lastly, there is notable variation between surgeon specific model accuracies. We suspect this may be due to the fact that even in a large dataset, there may be few individual surgeon-procedure combinations. The different procedures and human factors introduce a large amount of uncertainty. Institutions with more standardized cases may see even higher benefits of a machine learning approach. Future work includes a detailed analysis of what factors contribute to some surgeon specific models having higher accuracy than others.

Conclusion

The XGBoost ML surgeon-specific models had superior results compared to the ML random forest model, linear regression, and current standards of estimation. With the XGBoost surgeon-specific models, the ability to predict cases within the 10% tolerance threshold was improved in the testing dataset from 32% by the surgeon scheduler to 39%. The performance of the top performing individual surgeon models suggests that some individual surgeons may see predictions as high as 50% of cases falling within 10%. This is a significant improvement upon current standards of estimation.

Our study is a notable advancement towards statistical modeling of case-time duration across all surgical departments. We demonstrate the advantages of developing XGBoost machine learning models individualized per surgeon, and the potential efficiency improvements that can be achieved with this approach in a tertiary hospital. Our work suggests that machine learning models tailored to individual surgeons may help improve the management and scheduling of the operating room.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Abbreviations

ASA	American Society of Anesthesiologists
CPT	Current Procedural Terminology
EDA	Exploratory data analysis
EMR	Electronic medical record
ICD	International classification of diseases
MAPE	Mean absolute percent error
ML	Machine learning
OR	Operating room
PPiQSO	Center for Perioperative & Pain Initiatives in Quality Safety Outcome

SD	Standard deviation
UWMC	University of Washington Medical Center
XGBoost	Extreme Gradient Boosting

References

- Gillespie BM, Chaboyer W, Fairweather N. Factors that influence the expected length of operation: results of a prospective study. [Internet].BMJ Qual. Saf [Internet].; 2012;21:3–12. Available at: <http://qualitysafety.bmj.com/lookup/doi/10.1136/bmjqs-2011-000169>.
- Childers CP, Maggard-Gibbons M. Understanding Costs of Care in the Operating Room. [Internet].JAMA Surg [Internet].; 2018;90095:e176233 Available at: <http://archsurg.jamanetwork.com/article.aspx?doi=10.1001/jamasurg.2017.6233>.
- Macario A Are Your Hospital Operating Rooms “Efficient.” .Anesthesiology; 2006;105:233–234. [PubMed: 16871053]
- Stepaniak PS, Heij C, Mannaerts GHH, et al. Modeling procedure and surgical times for current procedural terminology-anesthesia-surgeon combinations and evaluation in terms of case-duration prediction and operating room efficiency: A multicenter study. .Anesth. Analg; 2009;109:1232–1245. [PubMed: 19762753]
- Laskin DM, Abubaker AO, Strauss RA. Accuracy of predicting the duration of a surgical operation. .J. Oral Maxillofac. Surg; 2013;71:446–447. [PubMed: 23351763]
- Roque DR, Robison K, Raker CA, et al. The accuracy of surgeons’ provided estimates for the duration of hysterectomies: A pilot study. .J. Minim. Invasive Gynecol; 2015;22:57–65. [PubMed: 25020086]
- May JH, Spangler WE, Strum DP, Vargas LG. Prod. Oper. Manag. 2011;20:392–405.
- Dexter F, Dexter EU, Masursky D, Nussmeier NA. Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. .Anesth. Analg; 2008;106:1232–1241. [PubMed: 18349199]
- Wu A, Brovman EY, Whang EE, et al. The Impact of Overestimations of Surgical Control Times Across Multiple Specialties on Medical Systems. .J. Med. Syst; 2016;40:1–6. [PubMed: 26573639]
- Master N, Zhou Z, Miller D, et al. Improving predictions of pediatric surgical durations with supervised learning. [Internet].Int. J. Data Sci. Anal [Internet]. Springer International Publishing; 2017;4:35–52. Available at: <http://link.springer.com/10.1007/s41060-017-0055-0>.
- Larsson A The accuracy of surgery time estimations. .Prod. Plan. Control; 2013;24:891–902.
- Zhou J, Dexter F, MacArio A, Lubarsky DA. Relying solely on historical surgical times to estimate accurately future surgical times is unlikely to reduce the average length of time cases finish late. .J. Clin. Anesth; 1999;11:601–605. [PubMed: 10624647]
- Zahoor A The role of Machine Learning in Predicting CABG Surgery Duration. [Internet].Proc. 27th Int. Conf. Mach. Learn. [Internet].; 2011:1–56. Available at: <http://www.icml2010.org/papers/903.pdf>.
- Van Veen-Berkx E, Elkhuizen SG, Van Logten S, et al. Enhancement opportunities in operating room utilization; With a statistical appendix. [Internet].J. Surg. Res [Internet]. Elsevier Inc; 2015;194:43–51. Available at: 10.1016/j.jss.2014.10.044. [PubMed: 25479906]
- Van Eijk RPA, Van Veen-Berkx E, Kazemier G, Eijkemans MJC. Effect of Individual Surgeons and Anesthesiologists on Operating Room Time. .Anesth. Analg; 2016;123:445–451. [PubMed: 27308953]
- Strum DP, Sampson AR, May JH, Vargas LG. Surgeon and Type of Anesthesia Predict Variability in Surgical Procedure Times. [Internet].Anesthesiology [Internet].; 2000;92:1454–1466. Available at: <http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00000542-200005000-00036>. [PubMed: 10781292]
- Pandit JJ, Tavare A. Using mean duration and variation of procedure times to plan a list of surgical operations to fit into the scheduled list time. [Internet].Eur. J. Anaesthesiol. [Internet].;

- 2011;28:493–501. Available at: <http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00003643-201107000-00006>. [PubMed: 21623186]
18. Paalvast M, Meeuwsen F C TDMJ. [273]Real-time estimation of surgical procedure duration. . 2015:6–10.
19. Ng N, Gabriel RA, McAuley J, et al. Predicting Surgery Duration with Neural Heteroscedastic Regression. [Internet]. 2017;68:1–12. Available at: <http://arxiv.org/abs/1702.05386>.
20. Olsen AB, Hu G, Wang L, Montabon F. Improvement of surgery duration estimation using statistical methods and analysis of scheduling policies using discrete event simulation. . 2015.
21. Hosseini N, Sir M, Jankowski C, Pasupathy K. Surgical Duration Estimation via Data Mining and Predictive Modeling: A Case Study. [Internet].AMIA ... Annu. Symp. proceedings. AMIA Symp. [Internet].; 2015;2015:640–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/26958199><http://www.ncbi.nlm.nih.gov/pubmed/26958199>.
22. Edelman ER, van Kuijk SMJ, Hamaekers AEW, et al. Improving the Prediction of Total Surgical Procedure Time Using Linear Regression Modeling. [Internet].Front. Med [Internet].; 2017;4:1–5. Available at: <http://journal.frontiersin.org/article/10.3389/fmed.2017.00085/full>.
23. Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for next-generation interventions. [Internet].Nat. Biomed. Eng [Internet]. Nature Publishing Group; 2017;1:691–696. Available at: <http://www.nature.com/articles/s41551-017-0132-7>. Accessed September 30, 2018. [PubMed: 31015666]

Figure 1: Training and Testing datasets

The dataset used for the testing and training of statistical models. The inclusion criteria were: weekday cases between January 1, 2014 and December 31, 2017, adult patients (age > 18), and procedure performed in main operating rooms. The machine learning and linear regression models were developed on the training dataset and validated on the testing dataset. The performance of the current methods of surgeon scheduler and EMR estimations was assessed on the testing dataset (N=9,398) to enable comparison with the proposed models.

Figure 2: Distribution of Prediction Error in Testing Dataset

The error distributions of the predictions by the surgeon scheduler (blue), XGBoost surgeon-specific model (composite of the 92 individual models; red), the EMR estimate using the prior 10 surgeon-primary procedures (gold), and XGBoost service-specific (composite of the 12 service models; purple). The 0 bin reflects -10% to 0%. Positive error represents underestimation while negative error represents overestimation. The red box denotes the -10 to 10% tolerance threshold for within cases. The surgeon-specific model had the best predictions within 10% as illustrated by the highest frequency within 10% and narrower distribution. The surgeon scheduler tends to underestimate while the EMR average tends to overestimate the case duration. The service specific model has similar performance with less underestimation.

Figure 3: Distribution of Accuracies of Surgeon-Specific Models:

The range of accuracies within the 92 surgeon-specific XGBoost models. Models varied in their accuracy. Models with accuracy greater than or equal to that of schedulers (ie >75%) constituted 45% of all models. These models were notably superior than the surgeon schedulers with within 10% prediction as high as 50% compared to 32%.

Table 1:
Preoperative data parameters used for predictive models

Perioperative data used for model development classified by relationship to patient, procedure, or personnel.

Patient factors: • Age   • Sex   • BMI   • Patient admit class (Inpatient/Outpatient)   • Preoperative diagnosis (ICD codes)   • Medical history conditions
Procedure factors: • Primary procedure   • Primary procedure category   • First/Second/Third/Fourth/Fifth subprocedure(s)   • Surgery modifiers: Robot, Revision, Laser, Laparoscopic, etc.
Personnel factors: • Surgeon unique identifier   • Historical primary procedure duration (at a surgeon level)   • Historical subprocedure duration (at a surgeon level)

Table 2:
Predicted case-time duration and outcomes for all models

The first two rows illustrate the results for the surgeon scheduler standard and the EMR estimates. The remaining rows indicate the linear regression and machine learning results.

Table 2A		Comparison of <u>Surgeon Specific</u> Models to Surgeon Scheduler				
Model	R ²	MAPE	Accuracy±SD	Overage	Underage	Within
Surgeon Scheduler (N=9,398)		25%	75±27%	39%	29%	32%
Average of last 10 Procedures (EMR Default) (N=9,242)		30%	70±42%	30%	40%	30%
Linear (N=7,854)	57%	36%	64±45%	33%	41%	26%
Random Forest (N=7,854)	91%	28%	72±36%	29%	40%	31%
XGBoost (N=7,854)	85%	26%	74±35%	34%	27%	39%
Table 2B		Comparison of <u>Service Specific</u> Models to Surgeon Scheduler				
Model	R ²	MAPE	Accuracy±SD	Overage	Underage	Within
Surgeon Scheduler (N=9,398)		25%	75±27%	39%	29%	32%
Average of last 10 Procedures (EMR Default) (N=9,242)		30%	70±42%	30%	40%	30%
Linear (N=9398)	55%	39%	61±51%	33%	44%	23%
Random Forest (N=9398)	93%	29%	71±38%	33%	44%	23%
XGBoost (N=9398)	77%	27%	73±34%	39%	29%	32%

MAPE = Mean absolute percentage error

SD = Standard deviation

Overage = % cases with actual case time duration > predicted + 10% tolerance threshold

Underage = % cases with actual case-time duration < predicted - 10% tolerance threshold

Within = % cases with actual case-time duration within ± 10% tolerance threshold

Table 3:
Main features used by surgeon-specific machine learning models (N=92) to predict case-time duration.

The top predictor variables utilized in the 92 surgeon-specific models. To determine the most impactful features overall, the weighted importance was calculated by multiplying the percentage gain of each feature by their percentage frequency of occurrence in the models. The relative contributions of procedure, surgeon and patient specific features on model predictions are shown in the first row.

Relative contributions of features to model predictions		
Description	Weighted feature gain	Feature Category
Average case-time duration of latest ten surgeries at Procedure level	25.1%	Procedure
Average case-time duration of latest ten surgeries at Surgeon and Procedure level	23.6%	Surgeon
Average case-time duration of latest ten surgeries at First SubProcedure level	11.3%	Procedure
Average case-time duration of latest ten Surgeries at Surgeon and First SubProcedure level	9.1%	Surgeon
Inpatient class	5.4%	Patient
Average case-time duration of latest ten surgeries at Surgeon level	3.1%	Surgeon
Age of the Patient	3.0%	Patient
Body Mass Index (BMI)	2.8%	Patient
Number of SubProcedures	2.4%	Procedure
Average case-time duration of latest ten surgeries at Second SubProcedure level	1.8%	Procedure
Number of preoperative problems patient's medical history	1.2%	Patient
Average case-time duration of latest ten surgeries at Third SubProcedure level	1.0%	Procedure
Average case-time duration of latest ten Surgeries at Surgeon and Second SubProcedure level	0.9%	Surgeon
Number of admission ICD codes	0.9%	Patient
Robotic procedures	0.7%	Procedure
ICD: Neoplasms	0.5%	Patient
Laparoscopic procedures	0.3%	Procedure
ICD: Diseases of the circulatory system	0.1%	Patient
Procedures using Laser	0.1%	Procedure
Male gender	0.1%	Patient
ICD: Diseases of the digestive system	0.1%	Patient
Medical history: CANCER	0.1%	Patient
ICD: Diseases of the nervous system	0.1%	Patient
Medical history: ARRHYTHMIA	0.1%	Patient
Medical history: ENDOCRINE (DIABETES)	0.1%	Patient
Medical history: SMOKING HISTORY	0.1%	Patient
Medical history: COAGULOPATHY	0.1%	Patient