Improving Operating Room Efficiency: Machine Learning Approach to Predict Case-Time Duration

Matthew A Bartek, MD, MPH, Rajeev C Saxena, MD, MBA, Stuart Solomon, MD, Christine T Fong, MS, Lakshmana D Behara, MS, Ravitheja Venigandla, MS, Kalyani Velagapudi, PhD, John D Lang, MD, Bala G Nair, PhD

BACKGROUND: Accurate estimation of operative case-time duration is critical for optimizing operating room use. Current estimates are inaccurate and earlier models include data not available at the time of scheduling. Our objective was to develop statistical models in a large retrospective data set 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. For each of these models, we generated an all-inclusive model, service-specific models, and surgeon-specific models. In the latter 2 approaches, individual models were created for each surgical service and surgeon, respectively. Our data set included 46,986 scheduled operations performed at a large academic medical 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 with our institutional standard of using average historic procedure times and surgeon estimates. 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 machine learning algorithm resulted in the highest predictive capability. The surgeon-specific model was superior to the service-specific model, with higher accuracy, 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 machine learning surgeon-specific model.

CONCLUSIONS: Our study is a notable advancement toward statistical modeling of case-time duration across all surgical departments in a large tertiary medical center. Machine learning approaches can improve case duration estimations, enabling improved operating room scheduling, efficiency, and reduced costs. (J Am Coll Surg 2019;229:346−354. © 2019 Published by Elsevier Inc. on behalf of the American College of Surgeons.)

The operating room (OR) is among the highest hospital revenue generators and accounts for as much as 42% of a hospital's revenue.[1,2] On the other hand, it also accounts for a high cost of use, estimated at $36 per minute.[1,2] Therefore, optimizing OR use is vital for delivering efficient and cost-effective care. A key first step toward scheduling surgical procedures is estimating their duration. Accurate estimation enables optimal case scheduling,

Drs Bartek and Saxena contributed equally to this work.

Disclosure Information: Nothing to disclose.

Disclosures outside the scope of this work: Dr Nair holds equity in and is a paid consultant to Perimatics, LLC.

Support: Dr Bartek was supported by a training grant from the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number T32DK070555.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Presented at the American College of Surgeons 104th Annual Clinical Congress, Scientific Forum, Boston, MA, October 2018, and the Washington/Oregon Surgical Society Annual Meeting, Sunriver, OR, June 2019.

Received December 9, 2018; Accepted May 30, 2019.

From the Departments of General Surgery (Bartek) and Anesthesiology and Pain Medicine (Saxena, Solomon, Fong, Lang, Nair), University of Washington, Seattle, and Perimatics LLC, Bellevue (Behara, Venigandla, Velagapudi), WA.

Correspondence address: Matthew A Bartek, MD, MPH, 3424 Meridian Ave N, Apt 10, Seattle, WA 98103. email: bartek@uw.edu

https://doi.org/10.1016/j.jamcollsurg.2019.05.029
ISSN 1072-7515/19

Abbreviations and Acronyms

EMR	= electronic medical record
ML	= machine learning
OR	= operating room
XGBoost	= Extreme Gradient Boosting

appropriate allocation of resources (equipment, personnel, and facilities), and creation of efficient patient flows.[3] Inaccurate case-time duration estimates 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.[5-8] 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 intraoperative findings can strongly influence case-time duration. In addition, there are multiple patient, anesthetic, and system factors that might not be considered in the surgeon estimation.[9,10] Alternatively, in many electronic medical record (EMR) scheduling systems, historic 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 preoperative data available on the case that is being performed.[11-13]

Estimating case-time duration at our institution is currently based on 2 parameters: the surgeon and the primary procedure. The EMR scheduling system takes the primary procedure to be scheduled and generates the average of the case times for the previous 10 procedures performed by the 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 EMR 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 estimate of case-time duration represents the current standard.

Earlier studies have sought to improve on estimated case-time duration, though no single approach has gained wide acceptance.[4,11,14–21] Statistical regression models have been used to predict case-time duration and assess the relative importance of input variables.[11,22] For example, Edelman and colleagues[23] were able to reduce the variation in total procedure time predictions in elective ophthalmology cases by 25% with a linear regression model that used surgeons' presurgical estimates, as well as the type of operation, type of anesthesia, American Society of Anesthesiologists class, and patient age. Master and colleagues[11] 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. However, the models were trained on only 10 operations within a single specialty, limiting their generalizability.[23]

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 OR suites. In addition, many of the models did not restrict the model inputs to preoperatively available information only, 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, the use of "big data" and modern data science methods, such as ML have gained increasing attention for their ability to predict perioperative events and aid in clinical decisions.[24,25] Estimating case duration is particularly well suited for ML approaches, given that the data sets are large, well annotated, and potentially capture the numerous factors that can influence case-time duration. We developed linear regression and ML models to predict OR case-time duration. The predictions from these models and from surgeon estimates (an institutional standard) were compared retrospectively with actual case times. We hypothesized that both linear regression and ML models would provide more accurate case-time duration predictions than our current standards of EMR-derived historic averages and surgeon estimates.

METHODS

Definitions and model context

In all of the models created, we sought to optimize the case-time duration of an operation and compared the predicted values from the model with the actual values in a testing data set. Case-time duration was defined as the total minutes from patient entry into the OR to room exit ("wheels in to wheels out"). This duration was selected due to its importance in OR scheduling and use, as opposed to selecting "surgical case time" or "anesthesia case time." Due to the lack of a standard in the literature and that our average institutional case duration is longer than 2 hours, we defined a priori a 10% clinical tolerance threshold by which to classify cases relative to the prediction model performance. Therefore, "overage" cases were those where the actual case-time duration exceeded the predicted time by 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 10% of the predicted value. For short duration procedures where predicted time was fewer than 100 minutes and the 10% tolerance was therefore fewer than 10 minutes, the threshold was used as 10 minutes. These categorizations provided a practical method to evaluate the clinical relevance of a model prediction.

The University of Washington Medical Center is a major tertiary care center with more than 30 ORs and 450 patient beds. The hospital serves the Greater Seattle area and has a wide catchment region, including 5 states. The case complexity and average case-time duration are high. During the year 2017, the mean case-time duration at University of Washington Medical Center was 3 hours and 13 minutes in 14,345 cases. Our institutional standard system for predicting case-time duration starts with an EMR-derived historic average based on a specific surgeon performing a specific operation. Surgeons can then override this historic average with their own estimates, and do so two-thirds of the time.

Data sources

After obtaining IRB approval (study 00005331), we used a perioperative EMR database to obtain OR metrics as well as patient information to develop our predictive models. Notably, several personnel and procedure variables are not available preoperatively. These include anesthesiologist, anesthetic plan, nursing staff, billing CPT codes, and intraoperative events. Although model performance can be enhanced with these data, they would not be relevant in prospective implementation. Therefore, we only used variables that were available preoperatively in model development. Data from 12 surgical service lines was included in the data set, including General, Cardiac, Thoracic, Vascular, Transplantation, Neurosurgery, Plastic, Orthopaedic, Gynecology, Urology, Otolaryngology, and Oral-Maxillofacial Surgery. Secure and de-identified preoperative data for model development were provided by the Center for Perioperative and Pain Initiatives in Quality Safety Outcome at the University of Washington.

The model-building approach and inclusion and exclusion criteria are shown in Figure 1. The starting data set was composed of preoperative data for 4 years from January 2014 to December 2017 that included scheduled operations performed on weekdays for adult patients (aged 18 years and older). Procedures performed at off-site locations, such as the endoscopy suite and radiology or cardiology procedure rooms, were not included. Operations with key missing data were excluded, as shown in Figure 1. This resulted in a data set of 38,880 procedures that was used for model development and validation. The data were randomly divided into a training data set (31,026 cases; 80% of data) for model development and a testing data set (7,854 cases; 20% of all the data) to confirm model performance. The EMR and surgeon estimates were also evaluated on the same testing data set 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 eTable 1.

Model development

A total of 6 models were created and compared with our earlier institutional standard system. We created linear regression and supervised ML models using the Extreme Gradient Boosting (XGBoost) algorithm (see eDocument 1).[26] The XGBoost was chosen because of its overall popularity within the data science community and through our own testing of several ML algorithms during model development (see eTable 2). The ML models developed were non-parametric ensemble models that combine multiple predictive techniques to produce the strongest predictive power without overfitting the data. In our initial development and testing, the XGBoost model yielded better predictions of case duration than random forest. Although both models are decision tree-based, we chose to use the XGBoost model because it was more computationally efficient and therefore better suited for wider, more real-time implementation. Model development was performed in R: A Language and Environment for Statistical Computing software.

We used 2 approaches to develop the final prognostic models. First, we generated surgical service-specific models, where each model included features based on information about patients, surgeons, and procedures as data inputs. Different models were developed for each of the 12 surgical service lines. Second, we generated a series of surgeon-specific models in which surgeons, rather than surgical specialty, were modeled individually. Only surgeons who performed ≥100 procedures in the training data set were selected for surgeon-specific models, and this selection was then applied to the inputs for the service-specific models to avoid a selection bias when comparing these approaches. Therefore, a total of 12 service-specific

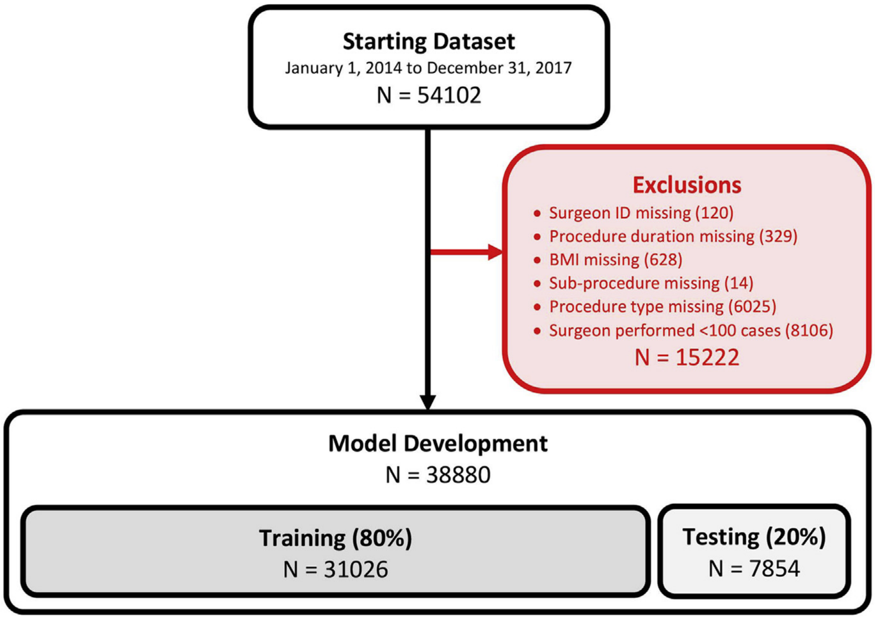
Starting Dataset
January 1, 2014 to December 31, 2017
N = 54102
Exclusions
• Surgeon ID missing (120)
• Procedure duration missing (329)
• BMI missing (628)
• Sub-procedure missing (14)
• Procedure type missing (6025)
• Surgeon performed <100 cases (8106)
N = 15222
Model Development
N = 38880
Training (80%)
N = 31026
Testing (20%)
N = 7854

Figure 1. Training and testing data sets. The data set 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 (aged older than 18 years), and procedure performed in main operating rooms. The machine learning and linear regression models were developed on the training data set and validated on the testing data set. The performance of the current methods of surgeon scheduler and electronic medical record estimations was assessed on the testing data set (n = 7,854) to enable comparison with the proposed models.

models corresponding to each surgical specialty and 93 surgeon-specific models for surgeons that met the minimum procedure count threshold were developed.

Table 1. Preoperative Data Parameters Used for Predictive Models

Parameter
Patient factor
Age
Sex
BMI
Patient admit class (inpatient/outpatient)
Preoperative diagnosis (ICD code)
Medical history condition
Procedure factor
Primary procedure
Primary procedure category
First/second/third/fourth/fifth subprocedure
Operative modifier, eg robot, revision, laser, laparoscopic
Personnel factor
Surgeon unique identifier
Historic primary procedure duration (at a surgeon level)
Historic subprocedure duration (at a surgeon level)

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

The model development followed a standard data science approach.[24] Categorical variables were converted into a binary representation for each category ("dummies"). For the linear regression models, given the right skewed distribution of all case-time durations, we applied a log transformation to the dependent variable (ie case-time duration). The selection of predictor variables for model development was based on those variables that passed the multicollinearity test (variance inflation factor <2.0). We then used both forward and backward stepwise regression to select independent variables to be included in the final models to optimize a given model's Akaike information criterion. We used a threshold p value of 0.05 for inclusion into the final linear model. For the non-parametric models, all features were used and the algorithm itself determined which remained in the models. All model parameters were determined based on the training data set.

Model evaluation

Models were used to predict case-time duration for procedures in the testing data set (20% of original) and the predictions were compared against the actual case-time duration. The 93 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 our institutional standards, the surgeon, and EMR estimations. The key metrics to measure model performance included the following: prediction accuracy: 1 minus mean absolute percent error, where mean absolute percent error is the mean absolute percent error; percentage overage: percentage of procedures that had overage; percentage underage: percentage of procedures that had underage; and percentage within: percentage number of cases with actual case-time duration within 10% of prediction (desired target metric).

All 6 models, differentiated by model type (linear or XGBoost) and approach type (all-inclusive, surgical service-specific, or surgeon-specific) were compared. For the service-specific and surgeon-specific models, the case-time duration for each operation was predicted using the corresponding service-specific or surgeon-specific model and these predictions were analyzed together.

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

Historic averages with surgeon estimate overrides accurately predicted case-time duration within a 10% tolerance threshold in 31% of cases, that is, 31% were "within," 42% of cases were overage cases (predictions underestimated the duration), and 27% of cases were underage cases (predictions overestimated the duration). The performance comparison of the proposed models and this institutional standard 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 93 individual models) performed better than the service-specific one (composite of 12 service-specific models), and the all-inclusive model had higher accuracy, lower percentage overage and percentage underage, and higher percentage within values. Figure 2 shows histograms of the distribution in prediction error for the surgeon, EMR estimation, service-specific model, and surgeon-specific model, further illustrating the highest performance of the surgeon-specific models.

The most accurate surgeon-specific model, via XGBoost, could predict 50% of cases accurately with a 10% tolerance threshold. The least accurate models were no worse than the surgeon predictions. 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. Features related to the surgeon accounted for 43% gain, those related to the procedure type accounted for 37% gain, and those related to the patient accounted for 15% gain in the surgeon-specific models. The list of features is presented in Table 3, including the categorization of the feature mentioned earlier. The majority of the information used in the models was based on procedure and personnel data. The 4 variables with highest gain included the average case time over the earlier 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 with personnel or procedure factors in predicting case duration.

DISCUSSION

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

The surgeon-specific ML models provided superior predictions compared with 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 earlier work in the literature. Master and colleagues[11] improved predictions compared with surgeons' predictions and historic averages, and found that of all input variables, the primary surgeon was the most impactful to decreasing variation in the model. Similarly, Strum and colleagues[17] showed that compared with patient factors and intraoperative variables, such as the anesthesiology team, type of anesthesia,

Table 2. Predicted Case-Time Duration and Outcomes for All Models

Model	R^2, %	MAPE, %	Accuracy, %, mean ± SD	Overage,* %	Underage,[†] %	Within,[‡] %
Surgeon (n = 7,854)	—	25	75 ± 27	39	29	32
Average of last 10 procedures (EMR default) (n = 7,854)	—	30	70 ± 42	30	40	30
All-inclusive model						
Linear (n = 7,854)	49	45	55	35	46	20
XGBoost (n = 7,854)	74	28	22	36	33	31
Service-specific model						
Linear (n = 7,854)	55	39	61 ± 51	33	44	23
XGBoost (n = 7,854)	77	27	73 ± 34	39	29	32
Surgeon-specific model						
Linear (n = 7,854)	57	36	64 ± 45	33	41	26
XGBoost (n = 7,854)	85	26	74 ± 35	34	27	39

The first 2 rows illustrate the results for the standard EMR historic average estimates and surgeon override estimates. The remaining rows indicate the linear regression and machine learning results. The SD denotes the distribution of individual model accuracies for service-specific and surgeon-specific ensembles. The "all-inclusive" models did not have associated SD because there was only one model used with this approach.

*Percent of cases with actual case-time duration > predicted + 10% tolerance threshold.

[†]Percent of cases with actual case-time duration < predicted − 10% tolerance threshold.

[‡]Percent of cases with actual case-time duration within ± 10% tolerance threshold.

EMR, electronic medical record; MAPE, mean absolute percentage error; XGBoost, Extreme Gradient Boosting.

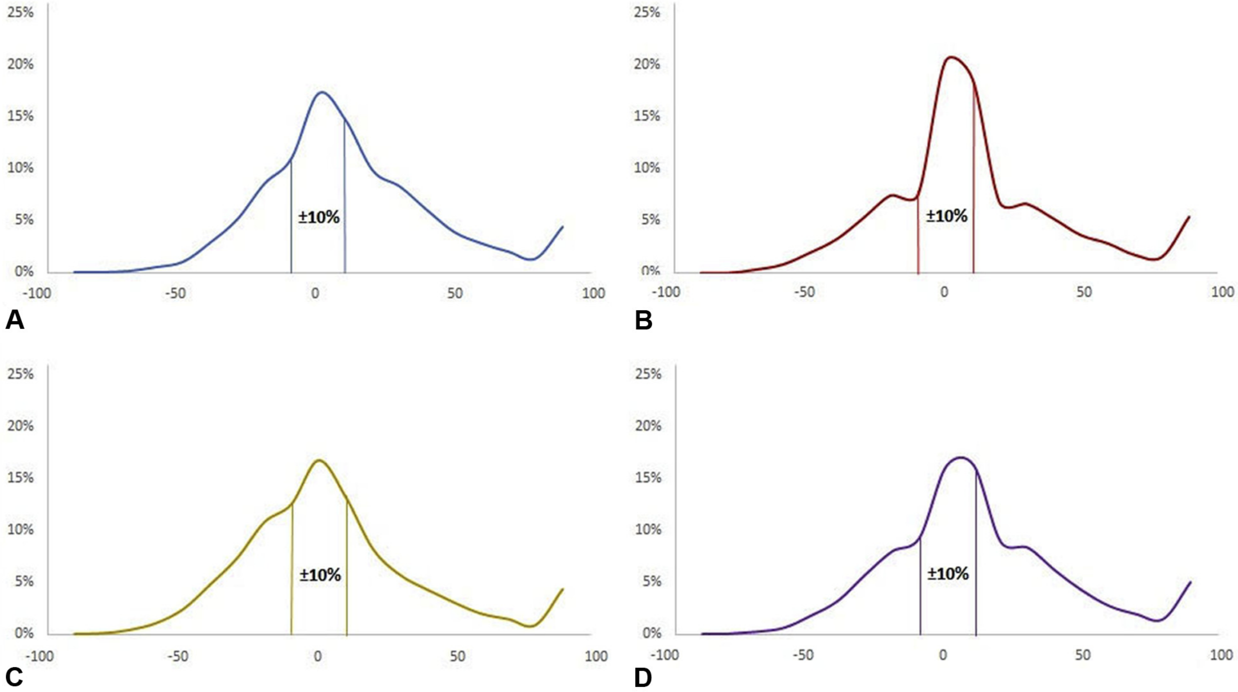

Figure 2. Distribution of prediction error in testing data set. (A) The error distributions of the predictions by the surgeon scheduler (blue), (B) Extreme Gradient Boosting (XGBoost) surgeon-specific model (composite of the 92 individual models; red), (C) the electronic medical record (EMR) estimate using the earlier 10 surgeon-primary procedures (gold), and (D) XGBoost service-specific (composite of the 12 service models; purple). The 0 bin reflects −10% to 0%. Positive error represents underestimation and 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 and the EMR average tends to overestimate the case duration. The service-specific model has similar performance with less underestimation.

Table 3. Main Features Used by Surgeon-Specific Machine Learning Models (n = 93) to Predict Case-Time Duration

Description	Weighted feature gain, %*	Feature category
Average case-time duration of latest 10 operations at procedure level	25.1	Procedure
Average case-time duration of latest 10 operations at surgeon and procedure level	23.6	Surgeon
Average case-time duration of latest 10 operations at first sub-procedure level	11.3	Procedure
Average case-time duration of latest 10 operations at surgeon and first sub-procedure level	9.1	Surgeon
Inpatient class	5.4	Patient
Average case-time duration of latest 10 operations at surgeon level	3.1	Surgeon
Age of the patient	3.0	Patient
BMI	2.8	Patient
No. of sub-procedures	2.4	Procedure
Average case-time duration of latest 10 operations at second sub-procedure level	1.8	Procedure
No. of preoperative problems patient's medical history	1.2	Patient
Average case-time duration of latest 10 operations at third sub-procedure level	1.0	Procedure
Average case-time duration of latest 10 operations at surgeon and second sub-procedure level	0.9	Surgeon
No. of admission ICD codes	0.9	Patient
Robotic procedure	0.7	Procedure
ICD: neoplasm	0.5	Patient
Laparoscopic procedure	0.3	Procedure
ICD: disease of the circulatory system	0.1	Patient
Procedure using laser	0.1	Procedure
Male sex	0.1	Patient
ICD: disease of the digestive system	0.1	Patient
Medical history: cancer	0.1	Patient
ICD: disease 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
ICD: disease of the musculoskeletal system and connective tissue	0.1	Patient
ICD: pregnancy, childbirth, and the puerperium	0.1	Patient

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.

*Weighted feature gain is an output of the Extreme Gradient Boosting algorithm.

and procedure code, surgeons are the most important source of variability in case-time duration predictions.

Non-ML, linear regression models performed poorly and were inferior to the surgeon and EMR estimations. We suspect this is because case estimation is not a linear problem and the assumptions concerning data characteristics for linear model development might not be valid. Surgeons 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 ±10% range, with fewer overage and underage cases compared with the surgeons. However, not all surgeon-specific models are equally accurate.

Despite surgeons overriding the EMR historic average estimate for two-thirds of scheduled cases, their ability to predict case duration within 10% was essentially the same, 32% for surgeon vs 30% for EMR. 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 used 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.[19] In cases where an unexpected bleed is encountered, it can be catalogued by the surgical staff and an "updated" time estimate could be generated. There can 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 can differ at community centers and other hospital settings. Although the models can improve estimation of case duration, this does not necessarily mean that OR use 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. To this end, each of the prediction models had a small peak of 4%—5% of cases in the 90% prediction error bracket (as seen in Fig. 2), which could affect scheduling significantly. Nonetheless, given that this proportion was comparatively equal among all of the modeling techniques, we believe this limitation does not substantially distinguish between model types. Lastly, there is notable variation between surgeon-specific model accuracies. We suspect this might be due to the fact that even in a large data set, there might be few individual surgeon-procedure combinations. The different procedures and human factors introduce a large amount of uncertainty. Institutions with more standardized cases might see even higher benefits of an ML approach. Future work includes a detailed analysis of what factors contribute to some surgeon-specific models having higher accuracy than others.

CONCLUSIONS

The XGBoost ML surgeon-specific models had superior results compared with linear regression, and current standards of case-time duration estimation, including an historic average by procedure type and surgeon and estimates provides by surgeons themselves. With the XGBoost surgeon-specific models, the ability to predict cases within the 10% tolerance threshold was improved in the testing data set from 32% by the surgeon to 39%. The performance of the top-performing individual surgeon models suggests that some individual surgeons might see predictions as high as 50% of cases falling within 10%. This is a significant improvement on current standards of estimation.

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

Author Contributions

Study conception and design: Bartek, Saxena, Solomon, Fong, Lang, Nair

Acquisition of data: Fong, Lang, Nair

Analysis and interpretation of data: Bartek, Saxena, Solomon, Fong, Behara, Venigandla, Velagapudi, Lang, Nair

Drafting of manuscript: Bartek, Saxena, Nair

Critical revision: Bartek, Saxena, Solomon, Fong, Behara, Venigandla, Velagapudi, Lang, Nair

REFERENCES

1. Gillespie BM, Chaboyer W, Fairweather N. Factors that influence the expected length of operation: results of a prospective study. BMJ Qual Saf 2012;21:3—12. Available at: http://qualitysafety.bmj.com/lookup/doi/10.1136/bmjqs-2011-000169. Accessed June 14, 2018.
2. Childers CP, Maggard-Gibbons M. Understanding costs of care in the operating room. JAMA Surg 2018;90095:e176233. Available at: http://archsurg.jamanetwork.com/article.aspx?doi=10.1001/jamasurg.2017.6233. Accessed June 14, 2018.
3. Macario A. Are your hospital operating rooms "efficient. Anesthesiology 2006;105:233—234.
4. 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.
5. Laskin DM, Abubaker AO, Strauss RA. Accuracy of predicting the duration of a surgical operation. J Oral Maxillofac Surg 2013;71:446—447.
6. 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.
7. May JH, Spangler WE, Strum DP, Vargas LG. The surgical scheduling problem: Current research and future opportunities. Prod Oper Manag 2011;20:392—405.
8. Zhou Z, Miller D, Master N, et al. Detecting inaccurate predictions of pediatric surgical durations, IEEE/ACM DSAA 2016. 17—19 Oct 2016. 3rd IEEE International Conference on Data Science and Advanced Analytics. New York: The Institute of Electrical and Electronics Engineers; 2016: 452—457.
9. 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.

10. 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.
11. Master N, Zhou Z, Miller D, et al. Improving predictions of pediatric surgical durations with supervised learning. Int J Data Sci Anal 2017;4:35—52. Available at: http://link.springer.com/10.1007/s41060-017-0055-0. Accessed June 14, 2018.
12. Larsson A. The accuracy of surgery time estimations. Prod Plan Control 2013;24:891—902.
13. 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.
14. Zahoor A, The role of machine learning in predicting CABG surgery duration. Proceedings of the 27th International Conference on Machine Learning, Haifa, Isreal, June 21—24. Available at: 2010. http://www.icml2010.org/papers/903.pdf. Accessed June 14, 2018.
15. Van Veen-Berkx E, Elkhuizen SG, Van Logten S, et al. Enhancement opportunities in operating room utilization; with a statistical appendix. J Surg Res 2015;194:43—51.
16. 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.
17. Strum DP, Sampson AR, May JH, Vargas LG. Surgeon and type of anesthesia predict variability in surgical procedure times. Anesthesiology 2000;92:1454—1466. Available at: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00000542-200005000-00036. Accessed June 14, 2018.
18. Pandit JJ, Tavare A. Using mean duration and variation of procedure times to plan a list of surgical operations to fit into the scheduled list time. Eur J Anaesthesiol 2011;28:493—501. Available at: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00003643-201107000-00006. Accessed June 14, 2018.
19. Guedon A, Paalvast M, Meeuwsen F, et al. Real-time estimation of surgical procedure duration. In: 2015 17th International Conference on E-health Networking, Application & Services (HealthCom). Boston, MA: IEEE; 2015:6—10. Available at: https://doi.org/10.1109/HealthCom.2015.7454464. Accessed June 14, 2018.
20. Ng N, Gabriel RA, McAuley J, et al. Predicting surgery duration with neural heteroscedastic regression. Available at: http://arxiv.org/abs/1702.05386. Accessed June 14, 2018.
21. 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. Iowa State University; 2015.
22. Hosseini N, Sir M, Jankowski C, Pasupathy K. Surgical duration estimation via data mining and predictive modeling: a case study. AMIA Annu Symp Proc 2015:640—648. eCollection 2015.
23. Edelman ER, van Kuijk SMJ, Hamaekers AEW, et al. Improving the prediction of total surgical procedure time using linear regression modeling. Front Med 2017;4:1—5. Available at: http://journal.frontiersin.org/article/10.3389/fmed.2017.00085/full. Accessed June 14, 2018.
24. Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for next-generation interventions. Nat Biomed Eng 2017;1:691—696. Available at: http://www.nature.com/articles/s41551-017-0132-7. Accessed September 30, 2018.
25. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018;2:749—760. Available at: https://doi.org/10.1038/s41551-018-0304-0. Accessed June 14, 2018.
26. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '16, 42. New York: ACM Press; 2016:785—794. Available at: http://dl.acm.org/citation.cfm?doid=2939672.2939785. Accessed June 14, 2018.

eDocument 1.
A Brief Explanation of Extreme Gradient Boosting

Traditional methods of modeling have relied on creating assumptions about the relationships between explanatory variables and dependent variables. For example, one key criterion for applying linear regression is that the relationship between predictors and the dependent variable is linear at all values of the variables. Broadly speaking, ML techniques have been used to "learn" the relationship between explanatory variables and dependent variables, without applying these a priori assumptions.

Extreme Gradient Boosting, the ML algorithm used in this study, is used for supervised learning problems like this one, where there is a training data set that includes values for both explanatory variables and the dependent variables—in this case, case-time duration. This algorithm balances the "training loss"—that is, how well the model fits the training data—with the complexity of the model—which can assess the degree of overfitting. Extreme Gradient Boosting has been used in a variety of competitions and has been shown to perform well among the variety of ML algorithms available. For more detailed information, see Chen and Guestrin.[26]

eTable 1. Predictor Variable Dictionary Used by the Predictive Models

Predictor variable, admission ICD_count	Description, no. of admission ICD code
ICD_1	1 Certain infectious and parasitic disease
ICD_2	2 Neoplasm
ICD_3	3 Endocrine, nutritional, and metabolic disease
ICD_4	4 Mental and behavioral disorder
ICD_5	5 Disease of the nervous system
ICD_6	6 Disease of the circulatory system
ICD_7	7 Disease of the respiratory system
ICD_8	8 Disease of the digestive system
ICD_9	9 Disease of the skin and subcutaneous tissue
ICD_10	10 Disease of the musculoskeletal system and connective tissue
ICD_11	11 Disease of the genitourinary system
ICD_12	12 Pregnancy, childbirth, and the puerperium
ICD_13	13 Certain conditions originating in the perinatal period
ICD_14	14 Congenital malformation, deformation, and chromosomal abnormality
ICD_15	15 Symptom, sign, and abnormal clinical and laboratory finding, not elsewhere classified
ICD_16	16 Injury, poisoning, and certain other consequence of external cause
ICD_17	17 Code for special purpose
ICD_18	18 External cause of morbidity and mortality
ICD_19	19 Factor influencing health status and contact with health service
ICD_20	20 Disease of the blood and blood-forming organ and certain disorder involving the immune mechanism
ICD_21	21 Disease of the eye and adnexa
ICD_22	22 Disease of the ear and mastoid process
SurgnPTAvg	Average OR duration of latest 10 operations at surgeon and procedure level
PTAvg	Average OR duration of latest 10 operations at procedure level
SurgnAvg	Average OR duration of latest 10 operations at surgeon level
SPID1Average	Average OR duration of latest 10 operations at first subprocedure level
SPID2Average	Average OR duration of latest 10 operations at second subprocedure level
SPID3Average	Average OR duration of latest 10 operations at third subprocedure level
SPID4Average	Average OR duration of latest 10 operations at fourth subprocedure level
SPID5Average	Average OR duration of latest 10 operations at fifth subprocedure level
SurSPID1Average	Average OR duration of latest 10 operations at surgeon, SubProcID1 level
SurSPID2Average	Average OR duration of latest 10 operations at surgeon, SubProcID2 level

(Continued)

eTable 1. Continued

Predictor variable, admission ICD_count	**Description, no. of admission ICD code**
SurSPID3Average	Average OR duration of latest 10 operations at surgeon, SubProcID3 level
SurSPID4Average	Average OR duration of latest 10 operations at surgeon, SubProcID4 level
SurSPID5Average	Average OR duration of latest 10 operations at surgeon, SubProcID5 level
SPID_Count	No. of sub-procedures
flag_PatientClassIn	Patient class = "inpatient"
Age	Age of the patient
BMI	Body mass index
flag_SexMale	Sex="Male"
PreOperProb_Count	No. of preoperative problems in medical history
POP_PULMONARYISSUES	Preoperative problem=pulmonary issues
POP_AIRWAYMANAGEMENT	Preoperative problem=airway management
POP_HYPERTENSION	Preoperative problem=hypertension
POP_CORONARYDISEASE	Preoperative problem=coronary disease
POP_HEARTFAILURE	Preoperative problem=heart failure
POP_VALVEDISEASE	Preoperative problem=valve disease
POP_ARRHYTHMIA	Preoperative problem=arrhythmia
POP_CRMD	Preoperative problem=cardiac rhythm management device
POP_PVD	Preoperative problem=peripheral vascular disease
POP_RENALDYSFUNCTION	Preoperative problem=renal dysfunction
POP_DIABETES	Preoperative problem=diabetes
POP_INSULINPUMP	Preoperative problem=insulin pump
POP_ENDOCRINE	Preoperative problem=endocrine
POP_CANCER	Preoperative problem=cancer
POP_OBESITY	Preoperative problem=obesity
POP_MORBIDOBESITY	Preoperative problem=morbid obesity
POP_HEPATIC	Preoperative problem=hepatic
POP_GERD	Preoperative problem=GERD
POP_IIP	Preoperative problem=increased intracranial pressure
POP_DEVELOPMENTALDELAY	Preoperative problem=developmental delay
POP_NLD	Preoperative problem=neurological disorder
POP_NMD	Preoperative problem=neuromuscular disorder
POP_MSD	Preoperative problem=musculoskeletal disorder
POP_COAGULOPATHY	Preoperative problem=coagulopathy
POP_VASCULARACCESS	Preoperative problem=vascular access
POP_SMOKINGHISTORY	Preoperative problem=smoking history
POP_IVDA	Preoperative problem=IVDA
POP_ETOHABUSE	Preoperative problem=EtOH abuse
ProcType_REVISION	Procedure contains the word *revision*
ProcType_REDO	Procedure contains the word *redo*
ProcType_LASER	Procedure contains the word *laser*
ProcType_ROBOT	Procedure contains the word *robot*
ProcType_LAPAROSCOPIC	Procedure contains the word *laparoscopic*

GERD, gastroesophageal reflux disease; IVDA, IV drug abuse; OR, operating room.

eTable 2. Comparison of Machine Learning Approaches

Machine learning technique	R^2, %	Accuracy,* %	Overage, %	Underage, %	Within, %
XGBoost	82	73	36	33	31
Random forest	95	70	29	41	30
CART	64	64	30	46	24
Artificial neural network	0	18	33	56	11

Before choosing XGBoost as the machine learning algorithm, we tested several other machine-learning algorithms and model approaches, including random forest, CART, and artificial neural networks. Models were all made using the "all-inclusive" approach, whereby the model was created to predict case-time duration and individual surgeons as well as service lines were represented by independent variables. The accuracy and "within" metric were both best with XGBoost as noted.

*Accuracy was defined as 1 − mean absolute percent error.

CART, Classification and Regression Trees; XGBoost, Extreme Gradient Boosting.