

**Development, deployment, and implementation of a machine learning surgical case length prediction model and prospective evaluation**

Hamed Zaribafzadeh, MS, Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA; Department of Surgery, Duke University, Durham, North Carolina, USA.

Wendy L. Webster, MA, MBA, Department of Surgery, Duke University, Durham, North Carolina, USA.

Christopher J. Vail, PA-C, MMCi, Department of Surgery, Duke University, Durham, North Carolina, USA.

Thomas Daigle, BA, Duke Health Technology Solutions, Duke University Health System, Durham, North Carolina, USA.

Allan D. Kirk, MD, PhD, Department of Surgery, Duke University, Durham, North Carolina, USA.

Peter J. Allen, MD, Department of Surgery, Duke University, Durham, North Carolina, USA.

Ricardo Henao, PhD, Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA.

Daniel M. Buckland, MD, PhD, Department of Emergency Medicine, Duke University, Durham, North Carolina, USA; Department of Mechanical Engineering and Materials Science, Duke University, Durham, USA.

Corresponding author: Daniel M. Buckland; 2301 Erwin Road, DUMC Box 3096, Durham, North Carolina, USA, 27710; +1 (919) 660-6553; dan.buckland@duke.edu

Funding source: Operational funding from the Department of Surgery, Duke University Medical Center.

## ABSTRACT

**Objective:** Implement a machine learning model using only the restricted data available at case creation time to predict surgical case length for multiple services at different locations.

**Background:** The Operating Room (OR) is one of the most expensive resources in a health system, estimated to cost \$22-133 per minute and generate about 40% of the hospital revenue. Accurate prediction of surgical case length is necessary for efficient scheduling and cost-effective utilization of the OR and other resources.

**Methods:** We introduced a *similarity cascade* to capture the complexity of cases and surgeon influence on the case length and incorporated that into a gradient boosting machine learning model. The model loss function was customized to improve the balance between over- and under-prediction of the case length. A production pipeline was created to seamlessly deploy and implement the model across our institution.

**Results:** The prospective analysis showed that the model output was gradually adopted by the schedulers and outperformed the scheduler-predicted case length in Aug-Dec 2022. In 33,815 surgical cases across outpatient and inpatient platforms, the operational implementation predicted 11.2% fewer under-predicted cases and 5.9% more cases within 20% of the actual case length compared with the schedulers and only over-predicted 5.3% more. The model assisted schedulers to predict 3.4% more cases within 20% of the actual case length and 4.3% fewer under-predicted cases.

**Conclusions:** We created a unique framework that is being leveraged every day to predict surgical case length more accurately at case posting time and could be potentially utilized to deploy future machine learning models.

ACCEPTED

## INTRODUCTION

The operating room (OR) is one of the most expensive resources in hospitals and comprises a significant portion of surgical costs, which accounts for 30-40% of all healthcare expenditures in the US<sup>1, 2</sup>. The OR costs \$22 to \$133 and generates about 40% of hospital revenue<sup>1-3</sup>. It is estimated that about 60% of all admitted patients require surgery during their hospital stay<sup>4, 5</sup>. Therefore, it is imperative to improve OR efficiency to enhance patient flow, increase hospital revenue, and reduce operating costs. One of the first steps toward improvement in OR efficiency is a more accurate surgical case length estimate for scheduling. Accurate prediction of surgical case length could increase the utilization and efficiency of the OR, reduce patient and surgeon wait time, and release pre-surgical beds for Post Anesthesia Care Unit (PACU)<sup>4, 6</sup>. However, a high degree of variability in the patient, procedure, surgeon, and operational factors causes case length prediction to be very challenging.

Several performance indicators such as over- and under-time case length predictions have been proposed to evaluate OR scheduling<sup>7</sup>. Over- or under-predicting a case length means that the case was finished earlier or later than the expected time, respectively. Over-prediction decreases the OR efficiency and increases OR idle time. In contrast, under-prediction leads to the cancellation or rescheduling of cases and increases patient wait time as well as the cost of surgery due to personnel overtime costs. About 55-59% of the total OR cost is the direct expenses, of which wages and benefits account for two-thirds in a California hospital study<sup>8</sup>. Although the financial effect of an improved OR utilization is complicated, a basic study showed that saving 9.6 minutes per case in an institution with 10,904 annual cases could save \$3.7 million annually which could be realized as more scheduled cases and staff utilization<sup>9</sup>. In another study, it was shown that a 21% reduction in under-prediction could save \$469,000 in

overtime costs over 3 years<sup>10</sup>. Strömblad et. al. showed that reducing under-prediction and absolute error resulted in decreased patient wait time with no increase in surgeon wait time<sup>6</sup>.

Current Procedural Terminology (CPT) codes have been used for over 25 years to post or create surgical cases<sup>11</sup>. Most hospitals use surgeon-estimated and/or historical median case length to schedule surgical cases, which both have uncertainties and are inaccurate<sup>1, 4-6, 10, 12, 13</sup>. Some hospitals also use their Electronic Health Record (EHR) system-generated time, which is a summary (median or average) of past similar cases lengths based on the surgeon, platform, and combination of CPT codes and is shown to be unreliable due to preoperative data variations<sup>5, 6, 13</sup>. Many groups have incorporated patient, procedural, and operational factors to create machine learning models and predict surgical case length<sup>1, 4</sup>. It has been shown that surgeons are the most significant contributor to case length variability as well as prediction<sup>13</sup>. Ito et. al. reviewed several machine learning approaches in the literature that have been used to predict surgical case length among which the boosted decision tree models showed better performance<sup>2</sup>. To the best of our knowledge, there is only one report in the literature that describes the implementation of a surgical case length predictive model in a clinical trial for multiple surgical services using preoperative data<sup>6</sup>. However, that study incorporated more than 300 variables from structured and unstructured data up to one day before surgery, which may not be necessarily available at case posting time. To seamlessly deploy a case length predictive model and provide schedulers with the predicted value at scheduling time, it is crucial to incorporate only the data that is available at case posting time which could be months before the surgery.

Here, we report the development, deployment, and implementation of a gradient-boosted decision tree model to predict surgical case length for multiple services and locations within our institution using very limited data available at the time of surgical case posting. We also propose

an improved method upon the EHR-generated median case length to calculate the historical median of past similar cases length and show that it improved model performance and significantly reduces sparsity and training time. Finally, we show how to adjust the model's loss function to balance over- and under-time errors based on our institution's priority.

## **METHODS**

### **Source of Data**

This study was found exempt by the Duke University Institutional Review Board (protocol number: Pro00104275). Elective surgical case posting data of 20 different services across 12 inpatient and ambulatory locations at Duke University Health System (DUHS) from Jul 2013 to Apr 2022 was initially collected and used to evaluate the best training period (results not shown), and the data from Jan 2021 to Apr 2022 was selected for the study. Records with missing timestamps or negative case length were excluded from the cohort which resulted in 107,898 cases. We selected 80,595 and 27,303 cases performed in 2021 and 2022 as the training and testing sets, respectively.

### **Outcome**

The patient-in to patient-out or wheels-in to wheels-out time was defined as the surgical case length because of its wide perioperative and scheduling use<sup>1, 13</sup>. The surgical case length was log-transformed to address the skewness of the original case length distribution (Supplemental Digital Content, Figure 1, Supplemental Digital Content 1, <http://links.lww.com/SLA/E646>).

### **Predictors**

The numerical variables include age, number of panels, number of posted CPTs, EHR-generated median case length, and surgeon-estimated case length. The categorical variables comprise

gender, patient class, service, primary physician, primary CPT, primary anesthesia type, location, and laterality. We previously reported that all the CPTs could be converted to Relative Value Units (RVU) in a case length prediction model<sup>14</sup>. The RVU consists of three categories: physician work, practice expense, and professional liability. The physician work or simply the work RVU accounts for about half of the total RVU and generally depends on the required time to perform a procedure<sup>15</sup>. Here we implemented the same method and converted all posted CPTs to a single work RVU as a continuous variable.

We defined a *similarity cascade* to calculate the historical median and standard deviation of case length (Figure 1). Cases in the training set with selected similar features at each stage were used to calculate the median and standard deviation of case length and stored as four separate reference tables. Then for each case, we looked at the first reference table to find similar features respective historical median and standard deviation of case length. If no similar case was found in the first table, the next tables with fewer similar features were used to assign the historical median and standard deviation of case length. If no similar case was found in the last reference table, a null value was assigned as the median and standard deviation.

## Model Development and Evaluation

Three gradient-boosted decision tree regression models were created using XGBoost<sup>16</sup> (version 1.2.1) in Python (version 3.7.6) to predict the case length. The mean squared logarithmic error (MSLE) was selected as the loss function. All the hyperparameters optimization was performed by 5-fold cross-validation on the training set. In the first model, all the above variables except the historical median and standard deviation were incorporated into the model. The primary physicians were one-hot encoded and used in the model. In the second model, all the above features except the one-hot encoded physicians were used and historical median and standard

deviation were derived from the reference tables created by the *similarity cascade* approach as two additional features. In the third model, we used a similar feature set as the second model but modified the loss function. Specifically, we introduced two regularization parameters C1 and C2 from 0.9 to 1.1 into the MSLE loss function for short ( $\leq 30$  minutes) and long ( $> 30$  minutes) case durations, respectively, to differently penalize cases in each group. Then we used Optuna<sup>17</sup> to find the best C1 and C2 parameters to minimize MSLE as well as the difference between over- and under-prediction errors. Each model performance was benchmarked against scheduler-predicted and EHR-generated case length using mean squared logarithmic error (MSLE) as well as mean absolute error (MAE), and root mean squared error (RMSE) as the performance metrics. Bootstrapping and resampling of the test set were used to calculate the 95% confidence intervals of the metrics.

### **Model Deployment and Implementation**

To seamlessly implement and deploy our case length prediction model, we collaborated with multiple groups within DUHS to create a production pipeline. Analytics Center of Excellence (ACE) created an application programming interface (API) that supplies surgical case data. Duke Institute for Health Innovation (DIHI) and Duke Health Technology Solutions (DHTS) supplied the framework to run the model on Duke Kubernetes system. The EHR system developers embedded a new field in Epic® OpTime to show the model-predicted case length where schedulers could easily incorporate the predicted value and schedule cases one day after the surgeon posted the case. The production pipeline starts one day after cases are posted. Data extraction, transformation, and loading (ETL) is completed in the backend by 4 AM and then the API is updated with the latest posted cases data by 5 AM. The model is run at 6 AM and the predicted case length is written as an embedded field in the Epic® for schedulers use by 7 AM.

Each step is separated by 1 hour to provide a reasonable buffer time against an unforeseen delay in the earlier steps. We registered our model and pipeline with the Duke Algorithm-Based Clinical Decision Support (ABCDS)<sup>18</sup> committee, an internal oversight board within our health system. After several phases of silent evaluation during May-July 2022, the model was implemented and the schedulers were directed to use the model-predicted case length beginning on August 1<sup>st</sup>, 2022. To measure and compare ORs utilization during silent evaluation and implementation phases, we defined and calculated the following three metrics for each phase: case volume as number of cases per day; percent room utilization as sum of all patient-in to patient-out time divided by sum of all time available; and nurse availability as number of nurses per case. All the three metrics were calculated considering cases that were started and ended between 7:30 AM - 7 PM and 7:30 AM - 4:30 PM in inpatient and ambulatory ORs, respectively, as regular operating hours.

## RESULTS

### Schedulers Performance

We evaluated the performance of the schedulers in the whole cohort (Jan 2021 to Apr 2022). Over- and under-prediction errors were defined as the case length predicted more or less than 20% of the actual case length, respectively. As shown in Figure 2, schedulers under-predicted cases 2.6 times more than over-predicting the cases indicating significant overbook of the ORs, patient and surgeon wait time, and staff over-time payment. Only 44% of cases were predicted within 20% of the actual case length.

### Model Retrospective Performance

We created three gradient-boosted decision tree models using the 2021 and 2022 data as training and testing sets, respectively, and compared the schedulers, EHR system-generated median, and models performance. The first model was created with encoded primary physicians and comprised 736 features of which 643 were only the one-hot encoded physicians. It predicted 57% of cases within 20% of the actual case length, 16.4% under, and 26.6% over the error margin (Figure 3). To demonstrate the significance of the *similarity cascade*, 643 encodes physicians were replaced in the second model with only 2 features, the historical median and standard deviation of case length calculated using the *similarity cascade*. Interestingly, it showed almost similar performance to the first model and predicted 57.8% of cases within the 20% margin while 17.6% and 24.6% of cases were under- and over-predicted, respectively showing an overall imbalanced error (Figure 3). Our health system requirement before implementation was an outcome with balanced over- and under-prediction errors. For example, the second model over-prediction for short cases ( $\leq 30$  minutes) is 2.6 times more than under-prediction indicating more imbalanced error within that case length period (Figure 4A). Such short cases are generally ambulatory cases with high volume and over-predicting those cases may negatively affect the number of performed cases and hospital revenue. Therefore, we created the third model using the same features set as in the second model and adjusted the loss function to improve the overall performance as well as the balance between the prediction errors. This model predicted 58.7% of cases within 20% of the actual case length with an overall balanced error of 20.7% over- and 20.6% under-prediction, respectively (Figure 3). Compared with the scheduler, model 3 resulted in 485 and 53 less over-time hours for cases that ended past regular working hours in inpatient and ambulatory ORs, respectively. Model 3 also predicted inpatient and ambulatory cases with 37.4 and 19.6 minutes less median total error time per room per day, respectively. Importantly,

all three models outperformed the scheduler as well as the EHR system-generated historical median, with the third model beating the other two (Figure 3). As shown in Figure 4B, the third model provides a more balanced prediction error for short cases and a 2.9% improvement in cases within the 20% margin as well as the lowest MSLE value (see Supplemental Digital Content, Table 1, Supplemental Digital Content 2, <http://links.lww.com/SLA/E647>). Therefore, the third model was selected for implementation.

### **Model Prospective Performance**

We collected 33,815 surgical cases that were performed at selected DUHS locations from Aug to Dec 2022 and evaluated the accuracy of model prediction as well as how much the model output was used by schedulers. Although the model generates predicted case length for all cases, schedulers were free to use and adjust the predicted value. Ideally, there should be no difference between schedulers and model case length prediction if the schedulers were exclusively using the model output for all cases. As shown in Figure 5A, the median difference between the schedulers and the model predicted case length value has been shrinking since the start of the silent evaluation in May 2022 when some schedulers adopted the use of model output and that gap has even further decreased after the model implementation in Aug 2022 and reduced by 7 minutes by the end of 2022 indicating that more schedulers gradually started to use the model output from Aug to Dec 2022. Schedulers performance was improved in Aug-Dec 2022 by using the model compared to Jan-Apr 2022 as they predicted 3.4% more cases within 20% of the actual case length and 4.3% fewer cases with under-time error and only over-predicted cases 1% more (Figure 3 and 5B). The model outperformed schedulers in all performance metrics (MSLE, MAE, and RMSE; see Supplemental Digital Content, Table 1, Supplemental Digital Content 2, <http://links.lww.com/SLA/E647>), performed more balanced, and predicted 11.2% fewer under-

predicted cases and 5.9% more cases within 20% of the actual case length compared with the schedulers and only over-predicted 5.3% more (Figure 5B), which the same performance trend can be seen at different predicted periods (Figure 5C and 5D). Specifically, the model outperformed the schedulers for short cases (i.e.,  $\leq 30$  minutes), 34.4% fewer under-predicted cases, 18.8% more cases within the 20% error margin, and only 15.6% more over-predicted cases. The implemented model also resulted in 5 and 26 fewer over-time hours for cases that ended past normal working hours as well as 18 and 20 minutes less median total error time per room per day in inpatient and ambulatory venues, respectively. We compared utilization metrics during silent evaluation and implementation phases, i.e., May-Jul 2022 and Aug-Dec 2022, respectively (see Supplemental Digital Content, Table 2, Supplemental Digital Content 3, <http://links.lww.com/SLA/E648>). The case volume decreased by 1.24 and increased by 2.14 cases per day in ambulatory and inpatient ORs, respectively, after the implementation. While there were 0.05 and 0.13 fewer nurses available per case in ambulatory and inpatient ORs, respectively, after the implementation due to our institution's nursing shortage, the utilization only decreased by 0.36% and 0.18% in ambulatory and inpatient venues, respectively.

## DISCUSSION

The *similarity cascade* addresses two main issues that could be raised by using a large number of encoded physicians in a model. First, a large number of encoded features creates a big sparse matrix which negatively affects computational cost. And second, the model is not robust to cases with new unseen primary physicians and fails. Replacing the encoded physicians in the second model with the historical median and standard deviation of case length calculated using the *similarity cascade* not only captured surgeon- and procedure-related case length variation in

various services and platforms but also resulted in significantly fewer features, improved robustness to new physicians, and approximately 5 times faster training than the first model.

It should be noted that the small amount of reduction in the MAE could significantly improve clinical workflow and save costs over time. For example, the fewer over-time hours for cases that ended past regular working hours in inpatient and ambulatory ORs could potentially reduce over-time labor expenses by about \$79,000 in Jan-Apr 2022 assuming 2 staffs per case for an over-time rate of \$73 per hour. Although model 3 adds some over-prediction or idle time, it would be possible to add more cases. For example, more than one thousand inpatient and ambulatory cases with lengths shorter than 20 and 40 minutes, respectively, were performed in Jan-Apr 2022 and similar short cases could be scheduled and performed using the saved time per room per day in inpatient and ambulatory ORs.

Prospective analysis of the implemented model showed gradual adoption of the model by schedulers. Specifically, the reduced difference in over-time error for cases that ended past normal working hours indicates an overall consistency between the model and scheduler for such cases. However, the model still outperformed the scheduler by less median total error time per room per day for both ambulatory and inpatient venues which implies that the unused OR time, especially the ambulatory rooms, could be better utilized by exclusively using the model for scheduling. Our institution experienced a historical nursing shortage in 2022, including surgical staff. Duke Surgery leadership directed the schedulers to use the model's predicted case length to avoid further disruption in everyday ORs workflow and maintain consistent scheduling. Although OR utilization is a complex topic that depends on several other factors, we managed to maintain the room utilization consistency in part by using the model during this national nursing shortage.

Our model has several limitations. First, new surgeons, CPTs, instruments, and techniques as well as increased efficiency in performing procedures could all affect the case length prediction. Although our model is robust to new physicians and CPTs because of the *similarity cascade* and use of RVUs it does not require new physicians to be encoded and explicitly added into the model. However, the model requires regular retraining and evaluation with the latest data to implicitly incorporate the above changes using the *similarity cascade*. Second, inconsistency in posting the CPTs could also negatively affect the case length prediction. Therefore, part of the implementation plan involved prompting the surgeons and their assistants to be as accurate as possible about CPTs when posting planned surgical cases. We also compared posted CPTs vs. billed CPTs and created lists of CPT combinations for each surgeon based on their posting vs. billed CPTs history to facilitate case creation and improve case posting consistency, so some of the improvement described above could simply have come from improved CPT accuracy in the case posting. Third, it has been shown that surgeon-estimated case length for cases with limited historical data could contribute to the case length prediction<sup>2, 11</sup>. Although surgeon-estimated case length is included in the implemented model not all surgeons input their estimate at case posting time. Therefore, we requested surgeons to provide their case length estimates, especially for new types of surgeries and/or complicated cases so that future updates could potentially improve the model predictions. Fourth, CPT codes are updated annually by the American Medical Association (AMA) and go into effect on Jan 1<sup>st</sup> of each year<sup>19</sup>. This could create inconsistency for some cases and negatively affect case length prediction. Fifth, our model was developed using the data from a single health system. Although we hypothesize that using similar feature set and the *similarity cascade* in other health systems should perform comparably<sup>20</sup>, our model needs to be validated at other institutions. Sixth, the features set in the

current model is from case posting data. More studies could explore the extraction and addition of reliable features from unstructured data using natural language processing (NLP) and image analysis, as well as other structured EHR data captured before case creation such as comorbidities and past surgical encounters. Seventh, our utilization metrics showed consistent room utilization before and after the model implementation. However, a deeper study is required to evaluate the model impact on OR utilization and efficiency as well as hospital revenue.

## Conclusions

We created the *similarity cascade* to better calculate the historical median of surgical case length and capture case complexity and primary surgeon influence on the case length. We showed that replacing thousands of primary surgeon names with a single median calculated by using the *similarity cascade* not only reduced the model training time and sparsity and enhanced robustness but also delivered the same, if not better performance. The *similarity cascade* could be customized and explored for other outcomes such as hospital length of stay (LOS) that are dependent on a variety of variables. We adjusted the model based on our institution's priority and created a production pipeline to deploy and implement the model for everyday use of the schedulers. Our prospective evaluation predicted 11.2% fewer under-predicted cases and 5.9% more cases within 20% of the actual case length compared with the schedulers and only over-predicted 5.3% more. We maintained room utilization while experience nursing shortage by using the model. The developed pipeline could be leveraged to deploy future models across our and other health systems.

## Acknowledgments

We would like to acknowledge Suresh Balu, Marshall Nichols, and Matt Gardner from DIHI, Ryan Craig, Alejandro Trillo, Deepthi Krisnamaneni, Sanjay Ghosh, and Omar Sodeq from ACE, and Andrii Kuraska and Tim Crittenden from Duke clinical applications for their support to create the production pipeline.

ACCEPTED

## References

1. Jiao Y, Sharma A, Ben Abdallah A, et al. Probabilistic forecasting of surgical case duration using machine learning: model development and validation. *J Am Med Inform Assoc* 2020; 27(12):1885-1893.
2. Ito M, Hoshino K, Takashima R, et al. Does case-mix classification affect predictions? A machine learning algorithm for surgical duration estimation. *Healthcare Analytics* 2022; 2.
3. Bellini V, Guzzon M, Bigliardi B, et al. Artificial Intelligence: A New Tool in Operating Room Management. Role of Machine Learning Models in Operating Room Optimization. *J Med Syst* 2019; 44(1):20.
4. Zhao B, Waterman RS, Urman RD, et al. A Machine Learning Approach to Predicting Case Duration for Robot-Assisted Surgery. *J Med Syst* 2019; 43(2):32.
5. Tuwatananurak JP, Zadeh S, Xu X, et al. Machine Learning Can Improve Estimation of Surgical Case Duration: A Pilot Study. *J Med Syst* 2019; 43(3):44.
6. Strömblad CT, Baxter-King RG, Meisami A, et al. Effect of a Predictive Model on Planned Surgical Duration Accuracy, Patient Wait Time, and Use of Presurgical Resources. *JAMA Surgery* 2021; 156(4).
7. Rahimi I, Gandomi AH. A Comprehensive Review and Analysis of Operating Room and Surgery Scheduling. *Archives of Computational Methods in Engineering* 2020; 28(3):1667-1688.
8. Childers CP, Maggard-Gibbons M. Understanding Costs of Care in the Operating Room. *JAMA Surg* 2018; 153(4):e176233.

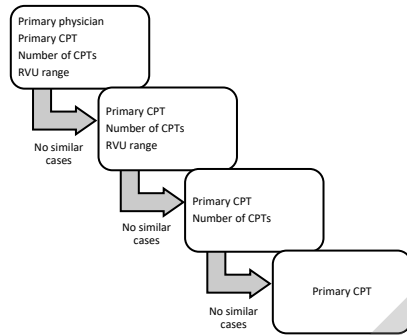
9. Miller LE, Goedicke W, Crowson MG, et al. Using Machine Learning to Predict Operating Room Case Duration: A Case Study in Otolaryngology. *Otolaryngol Head Neck Surg* 2022;1945998221076480.
10. Rozario N, Rozario D. Can machine learning optimize the efficiency of the operating room in the era of COVID-19? *Can J Surg* 2020; 63(6):E527-E529.
11. Dexter F, Epstein RH, Marian AA. Case duration prediction and estimating time remaining in ongoing cases. *Br J Anaesth* 2022; 128(5):751-755.
12. Robertson A, Kla K, Yagmour E. Efficiency in the operating room: optimizing patient throughput. *Int Anesthesiol Clin* 2021; 59(4):47-52.
13. Bartek MA, Saxena RC, Solomon S, et al. Improving Operating Room Efficiency: Machine Learning Approach to Predict Case-Time Duration. *J Am Coll Surg* 2019; 229(4):346-354 e3.
14. Garside N, Zaribafzadeh H, Henao R, et al. CPT to RVU conversion improves model performance in the prediction of surgical case length. *Sci Rep* 2021; 11(1):14169.
15. Nurok M, Gewertz B. Relative Value Units and the Measurement of Physician Performance. *JAMA* 2019; 322(12):1139-1140.
16. 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)*. Association for Computing Machinery, New York, NY, USA, 2016:785-794.
17. Akiba T, Sano S, Yanase T, et al. Optuna: A Next-generation Hyperparameter Optimization Framework. *In Proceedings of the 25th ACM SIGKDD International*

*Conference on Knowledge Discovery & Data Mining (KDD '19). Association for Computing Machinery, New York, NY, USA 2019:2623-2631.*

18. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. *J Am Med Inform Assoc* 2022; 29(9):1631-1636.
19. <https://www.ama-assn.org/about/cpt-editorial-panel/cpt-code-process>.
20. Lam SSW, Zaribafzadeh H, Ang BY, et al. Estimation of Surgery Durations Using Machine Learning Methods-A Cross-Country Multi-Site Collaborative Study. *Healthcare (Basel)* 2022; 10(7).

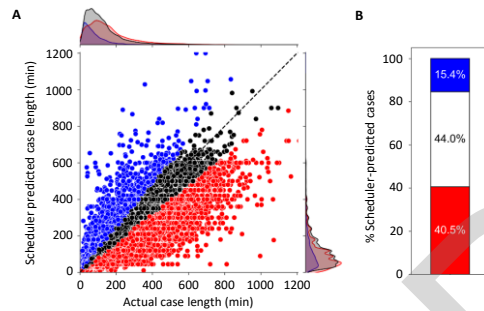
## Legends for Illustrations

**Figure 1.** The *Similarity cascade* to find similar surgical cases at each stage for calculation of the historical median and standard deviation of case length.

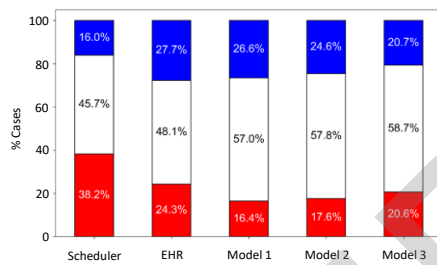


ACCEPTED

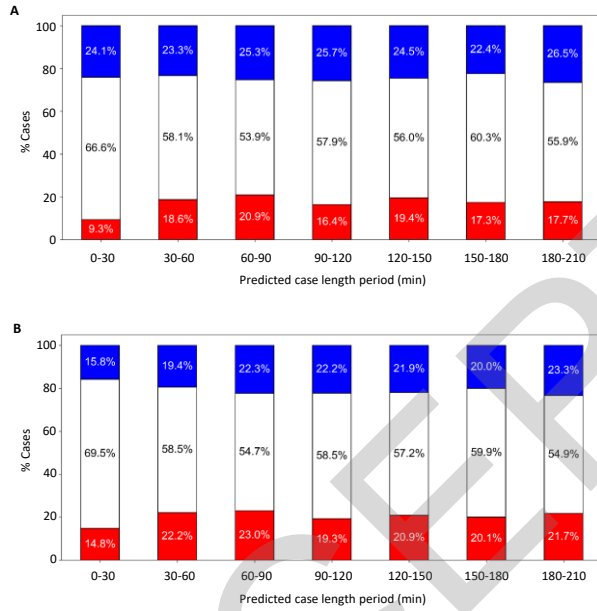
**Figure 2.** Schedulers' performance from Jan 2021 to Apr 2022. A) Distribution of the actual and the scheduler-predicted case length in minutes, and B) Percentage of cases predicted within, over, or under 20% of the actual case length. Predicted case lengths within, under, and over 20% of the actual case length are depicted in white, red, and blue, respectively.



**Figure 3.** Performance comparison of three models with scheduler and EHR predicted case length. Models 1-3 include encoded physicians, historical median from *similarity cascade*, and historical median from *similarity cascade* with adjusted loss function, respectively. Predicted case lengths within, under, and over 20% of the actual case length are depicted in white, red, and blue, respectively.



**Figure 4.** Performance comparison at different predicted case length periods. A) Model 2, and B) Model 3. Predicted case lengths within, under, and over 20% of the actual case length are depicted in white, red, and blue, respectively.



**Figure 5.** Prospective evaluation of the case length model from Aug-Dec 2022. A) Median difference in predicted case length per day by the scheduler and model in minutes, B) Overall performance comparison between scheduler and model, and C-D) scheduler and model performance at different predicted case periods, respectively. The solid black lines are the regression lines from May-July 2022 and Aug-Dec 2022. The dashed red and black lines indicate the implementation of the model from Aug 2022 and no difference between the median of the scheduler and model predicted case length per day, respectively. Predicted case lengths within, under, and over 20% of the actual case length are depicted in white, red, and blue, respectively.

