**Optimizing Surgical Efficiency: Predicting Case Duration in Common General Surgery Procedures Using Machine Learning**

Michelle Kwong, MD[[1]](#footnote-1),[[2]](#footnote-2)

Mohammad Noorchenarboo, MSc[[3]](#footnote-3)

Katarina Grolinger, PhD3

Jeff Hawel MD, FRCSC[[4]](#footnote-4)

Christopher M Schlachta MD, FRCSC4

Ahmad Elnahas, MD, FRCSC4,[[5]](#footnote-5)\*

Manuscript word count: \*\*\*

**KEY POINTS**

**Question:** Can machine learning models predict surgical case duration of adult elective general surgery procedures more accurately than provider knowledge-based estimates?

**Findings:** In this retrospective cohort study of 16,159 patients who had undergone 17,426 unique general surgery procedures, an artificial neural network model predicted surgical case duration over 18 minutes more accurately than the surgeon's estimate.

**Meaning:** Machine learning may be used to support providers to schedule cases objectively and thereby optimize operating room efficiency.

**ABSTRACT** 

**Importance:** Accurate prediction of surgical case time duration is critical to optimizing the use of operating room (OR) resources. Currently, surgical cases are scheduled according to subjective estimates by surgeons or staff, relying heavily on prior experiences or historical averages.

**Objective:** This study aims to develop and compare various prediction models, including machine learning-based algorithms and conventional statistical models, to objectively predict case time duration for common general surgical procedures.

**Main Outcome(s) and Measure(s):** Models were trained to predict “case time duration”, defined as the time between patient entry to and departure from the OR. Model performance was evaluated based on predictive accuracy, residual analysis, and benchmarked against “scheduled duration”, defined as case time estimated preoperatively by the primary surgeon and/or staff.

**Design:** Retrospective cohort study, using data from 2015 to 2020.

**Setting:** Multi-center, 3 academic tertiary institutions.

**Participants:** Adult patients undergoing elective general surgical procedures.

**Results:** Predictive models, including simple linear regression, Ridge regression, Lasso regression, Support Vector Regression, Random Forest, Gradient Boosting Machine, XGBoost, and Artificial Neural Network (ANN), were trained on a cohort of 16,159 patients [mean age, 56.85 ± 15.95; 47.48% male] having undergone 17,246 elective general surgery procedures. The ANN model demonstrated the superior predictive accuracy (Root Mean Squared Error, 49.7 minutes [95% CI 47.5 to 52.0]; Mean Absolute Error, 31.8 minutes [95% CI 30.6 to 33.0]). Residual analysis showed that the ANN resulted in an average residual of -0.37 minutes [95% CI -40.42 to 39.68, *p* = 0.34], while the scheduled duration produced an average residual of -18.52 minutes [95% CI -55.24 to 18.2, *p* < 0.01], demonstrating that the ANN provided a more accurate case time estimation by more than 18 minutes.

**Conclusions:** The ANN model provided more accurate case time estimates compared to provider knowledge-based estimates, while eliminating the subjective bias and dogma inherent in the traditional scheduling methods. Future research will explore whether the application of machine learning for the prediction of case time duration could improve healthcare resource utilization and costs.

**1 INTRODUCTION**

Optimizing operating room (OR) efficiency is vital to the delivery of high-quality and timely surgical care.1 To this end, accurate predictions of case duration are necessary to extract maximal value from limited OR resources.2 Inaccurate estimates lead to both under and over-utilization of OR time—resulting in idle time, overtime wages, frequent delays, case cancellations and postponements—that reduce surgical throughput, incur additional operational costs and, ultimately, impact patient care.1

Current industry standards for predicting case duration rely either on subjective surgeon estimates or on historical averages from electronic medical records (EMR), and often overlook patient, procedural, anesthetic and systemic factors that impact operative time3. These approaches have been shown to have limited accuracy, significant variability, and predictive inconsistency, highlighting the need for a more reliable method of estimation.3

Recent advancements in Machine Learning (ML) present promising alternatives to existing methods of case scheduling.4 ML has demonstrated potential in improving the accuracy of case duration predictions, which could enhance OR scheduling and decision-making processes.5,6 By integrating ML with EMR systems, healthcare providers can better manage surgical workflows and address the backlog, ultimately leading to more efficient and effective OR utilization.7

This study aims to develop advanced predictive models using a large dataset of common general surgery procedures from our institution’s EMR to enhance the accuracy of case-time duration estimates. We hypothesize that an ML model, incorporating a wide range of clinical variables beyond those considered in traditional methods such as patient characteristics, procedural details, and personnel information will offer more precise predictions.

# **2 METHODS**

This retrospective cohort study was approved by the Research Ethics Board of Western University (London, ON, Canada), and is here reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD-AI) guidelines.8

## **2.1 Data Sources & Setting**

Existing de-identified EMR (Cerner, North Kansas City, USA) data were retrospectively collected for surgical cases that took place across 3 tertiary care academic hospitals in the London Health Sciences Center (London, ON, Canada) over a 5-year period. All data was housed on a secure university-based server for analysis, model development, and internal validation.

## **2.2 Study Population**

From this dataset, a cohort of adult (age ≥ 18) patients undergoing general surgery procedures scheduled between January 1, 2015, and January 1, 2020, was identified. Included cases were limited to the following operations: appendectomy, non-radical cholecystectomy, colectomy, gastric bypass, non-hiatal abdominal hernia repair, ileostomy closure, liver resection, mastectomy, parathyroidectomy, thyroidectomy, and Whipple procedure. Urgent and emergent non-elective cases were excluded. Cancelled cases were also excluded. Cohort selection is outlined in Figure 1.

## **2.3 Data Preparation & Pre-processing**

Cases with missing values were removed rather than imputing missing values to avoid introducing bias. Outliers were identified based on clinical parameters (age ≥ 130; BMI ≤ 5, ≥ 200) and removed. Continuous features were scaled using min-max normalization. Categorical variables with rare categories (defined as those with frequency ≤ 1%) were merged into a generalized “Other” category. Then, all categorical variables underwent one-hot encoding or ordinal encoding. Additional features were derived from existing date-time data. To ensure transparency and reproducibility, a comprehensive description of all variables in the dataset and applied pre-processing steps are presented in Supplementary Table 1.

The final pre-processed dataset, consisting of 17,246 examples, was generated using structured query language (SQL; Joint Technical Committee of the International Organization for Standardization and International Electrotechnical Commission [ISO/IEC]) and Python programming language version 3.9 (Python Software Foundation) code.

### **2.4 Model Development**

The primary study outcome and predictive output was “case time duration”, defined as the time between patient entry to and departure from the OR. Models were designed to make predictions at the time of case scheduling by the primary surgeon; to prevent data leakage, all inputs were restricted to information available in the pre-operative period. Initial feature selection was informed by clinical field expertise as well as a review of the existing literature. Ten-fold cross validation was used for model training and testing, as illustrated in Supplementary Figure 2. The previously described data preparation and pre-processing steps were applied separately to each fold to prevent data leakage. The list of all variables used in this study is provided in Supplementary Table 1.

Traditional multivariate and machine learning supervised machine learning models were trained: simple linear regression, Ridge regression, Lasso regression, Support Vector Regression (SVR), Random Forest, Gradient Boosting Machine (GMB), XGBoost and Artificial Neural Network (ANN). For each model type, hyperparameters were tuned using the Tree-structure Parzen Estimator algorithm9, a variant of Bayesian optimization particularly effective in handling high-dimensional space, as illustrated in Supplementary Figure 4 and described in further detail in Supplementary Table 2.

### **2.5 Model Evaluation**

Predictive models were benchmarked against “scheduled duration”—the estimated case time booked by the primary surgeon. Model performance was evaluated via Mean Squared Error (MSE)10, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE)11, Mean Absolute Percentage Error (MAPE) and Coefficient of Determination (*R*2), calculated across each fold of cross-validation.12 MSE and RMSE both reflect the mean difference between predicted output and target variable while penalizing larger errors. MAPE and MAE evaluate accuracy.

**2.6 Statistical Analysis**

Statistical analysis was designed to compare models by testing for significant differences across performance metrics. Levene’s test was first conducted to confirm the assumption of equality of variances held true for each metric. To reduce the risk of Type I errors from performing multiple t-tests, Analysis of Variance (ANOVA) was used to compare groups of models against “scheduled duration”, and significant results were followed by Dunnett’s test post-hoc.

Analysis of residuals—the differences between actual and predicted outputs—was also performed to assess whether any models exhibited consistent bias. One-sample t-tests were performed to evaluate whether model residuals were significantly different from zero, with residuals not significantly different from zero indicating unbiased predictions.

**3 RESULTS**

**3.1 Descriptive Analysis**

A cohort of 16,159 patients having undergone 17,246 elective general surgery procedures were identified to train and internally validate predictive models. The most frequently scheduled cases were abdominal hernia repairs (33.92%), non-radical cholecystectomies (20.00%) and thyroidectomies (11.99%), followed by gastric bypasses (4.85%), liver resections (4.75%), parathyroidectomies (4.62%), mastectomies (1.86%) and appendectomies (1.07%). Cases were largely performed by the General Surgery service (83.35%). Additional patient and surgical case characteristics of the study population are described in Table 1.

**3.2 Model Performance**

The ANN consistently outperforms all other models, as shown by performance metrics in Table 2. With the lowest RMSE (49.7 ± 2.3) and MSE (2469.3 ± 224.1), it demonstrates superior predictive accuracy compared to XGBoost, which follows closely with an RMSE of 50.0 ± 2.2. The Scheduled Duration baseline also performs well with an RMSE of 49.9 ± 2.0, but traditional regression models like Ridge, Linear, and Lasso Regression show higher RMSE values, ranging from 55.9 to 56.0.

In terms of MAPE, both Neural Networks and Random Forest excel, with a low MAPE of 0.26 ± 0.01, indicating effective minimization of percentage errors. The Scheduled Duration baseline, however, shows a higher MAPE of 0.34 ± 0.01, which may reduce its predictive reliability compared to ML models. For R², the Neural Network achieves a value of 0.78 ± 0.01, matching the performance of XGBoost and the Scheduled Duration baseline, showing that all three models capture the variance in the data similarly well.

Levene's test confirmed the assumption of equal variances for each evaluation metric. Subsequent, ANOVA results revealed significant performance differences across the models for all metrics, with p-values below 0.05. The results of Dunnett's test, as shown in Table 3, indicate that for the metrics of MSE, RMSE, and R², the Scheduled Duration baseline outperforms traditional regression models (Linear, Ridge, and Lasso Regression). However, when comparing these metrics with more advanced models such as Neural Networks, Random Forest, XGBoost, and GBM, the performance is generally comparable to Scheduled Duration.

## **3.3 Residual Comparison**

As shown in Table 4, the ANN model is the only one that does not have a significant difference from zero (*p* = 0*.*337), suggesting that it provides more accurate predictions with less bias. In contrast, other models, including the Scheduled Duration, show significant differences from zero (*p <* 0*.*05), indicating potential biases in their predictions. When comparing the mean and median residuals, the Neural Network model demonstrates a mean residual close to zero (-0.37 minutes) and a median of -3.69 minutes, further underscoring its superior performance in minimizing prediction errors compared to other models, where the Scheduled Duration shows a larger mean residual of -18.52 minutes.

# **4 DISCUSSION**

This study presents compelling evidence that machine learning models, particularly Artificial Neural Networks (ANN), can provide more accurate predictions of case time duration than conventional methods. Traditional methods, predominantly reliant on surgeon intuition and historical averages, often result in substantial discrepancies between scheduled and actual surgical durations.9 By leveraging a comprehensive dataset of 17,246 elective surgical procedures, our findings indicate a clear advancement in predictive accuracy that could enhance OR management.

The superiority of the ANN model over other conventional regression models and machine learning approaches is noteworthy. The ANN model demonstrated a mean residual close to zero (-0.37 minutes) whereas the scheduled duration showed a larger mean residual of -18.52 minutes, resulting in a more accurate and unbiased case time estimation by over 18 minutes. This discrepancy underscores the limitations of relying on historical averages or personal experience for scheduling, which can lead to inefficient OR usage and extended patient wait times. Other previous studies have also demonstrated that machine learning models, particularly ANNs can provide superior predictive accuracy, reducing the mean absolute error significantly compared to conventional approaches.10

Before the development of ANN models, previous research utilized linear statistical models to evaluate the importance of relevant input variables and predict case time durations.11 Patient and clinical characteristics were the most utilized predictors of surgery duration, such as the algorithms developed to predict the duration of laparoscopic cholecystectomies.12 However, one important disadvantage with this strategy is the inability of these type of models to handle numerous input variables. Studies have found that consideration of temporal factors like surgical setting, rank of case, day, week or month of surgery should be considered to improve prediction.13,14

Machine learning algorithms leverage vast amounts of data and complex variables that encompass patient demographics, case complexity, and provider experience, enabling a more nuanced understanding of predictive factors.15 One of the most significant implications of our findings is the reduction of subjective bias associated with provider-based estimates. The ANN model’s mean residual was close to zero, indicating that its predictions were unbiased and reliable. This not only highlights the potential for improved accuracy but also the elimination of biases inherent in human judgment.16 This suggests that integrating machine learning into surgical scheduling could mitigate the variability inherent in surgeon or staff estimations. As surgical procedures become increasingly complex, employing data-driven models like the ANN could enhance the precision of time predictions, ultimately leading to better OR efficiency and patient throughput.17

There are some limitations that should be acknowledged in this study. First, the study’s reliance on retrospective data from a single academic institution may limit the generalizability of the results. Variability in surgical practices, patient populations, and institutional resources could impact the applicability of the model to other settings.18 Furthermore, the algorithm’s performance could be influenced by the quality of the input data, including missing values or inaccuracies within electronic medical records. Future studies should explore multi-center data validation to assess the robustness of machine learning models across diverse healthcare environments.19 Additionally, while machine learning models can improve prediction accuracy, they do not account for real-time intraoperative variables, such as unexpected complications or variations in surgical techniques. These factors can significantly influence case durations and may not be captured in preoperative data.20 Developing adaptive models that can incorporate real-time data could further enhance predictive capabilities and operational efficiency.

Clinically, the implications of implementing machine learning algorithms are profound. Improved accuracy in predicting surgical case time allows for better scheduling, optimizing the use of surgical suites and personnel.21 This could lead to reduced costs and enhanced patient satisfaction due to reduced cancellations and more efficient care delivery.22 Moreover, machine learning could facilitate the identification of cases at risk of extended durations, allowing for preemptive adjustments in scheduling and resource allocation. While this study focused on elective general surgical procedures, the framework developed here could be adapted to other surgical specialties, potentially leading to widespread improvements across the healthcare system.

**5 CONCLUSIONS**

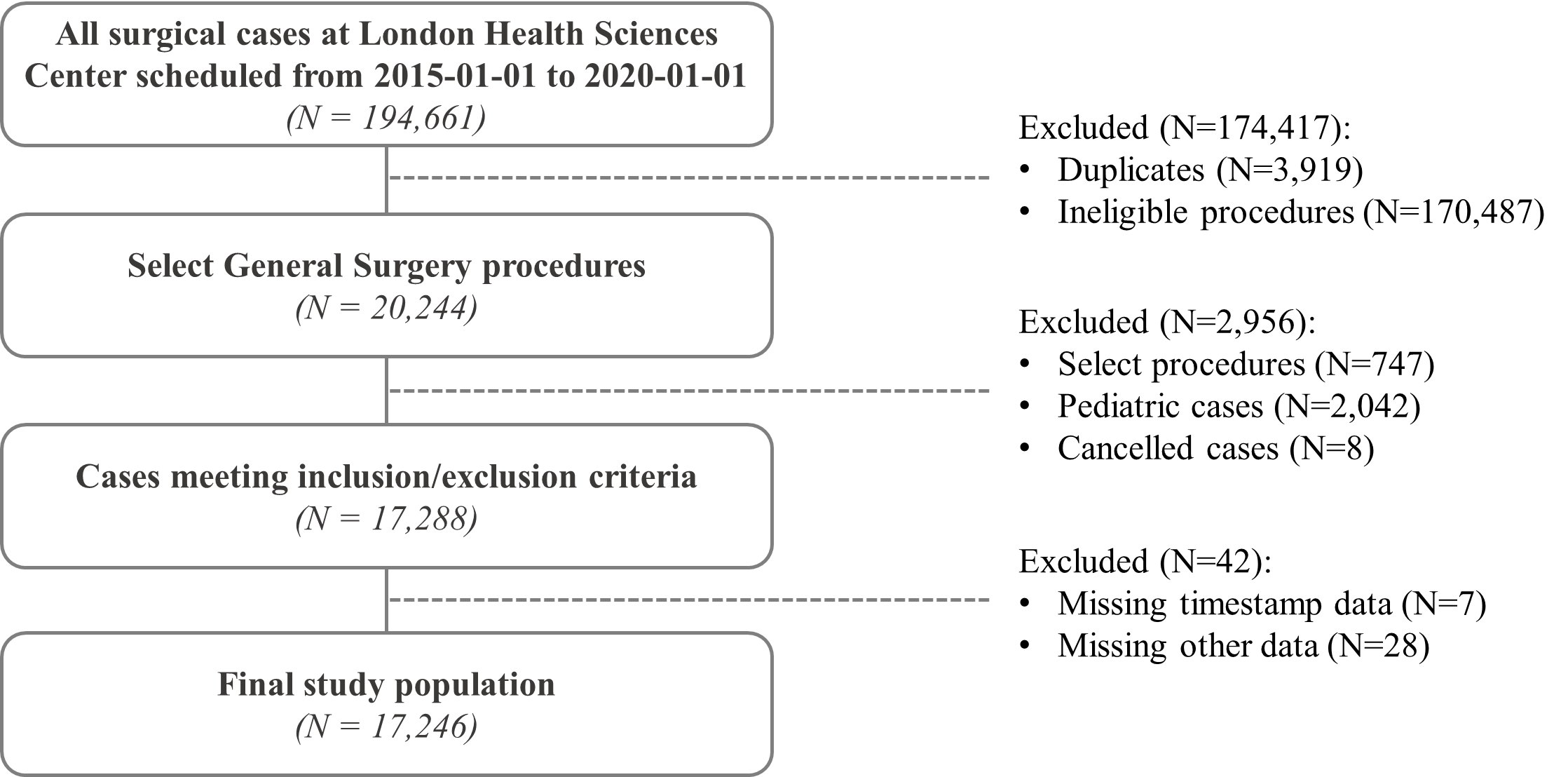
In conclusion, our study demonstrates that machine learning, specifically the ANN model for predicting surgical case time duration, provides a significant advantage in predicting surgical case durations compared to traditional estimation methods. By minimizing bias and improving accuracy, these advanced predictive models can optimize OR scheduling, potentially leading to more efficient use of healthcare resources. Continued exploration of this technology will be crucial in further refining surgical processes and improving overall patient outcomes. The next stage of research will explore whether integration of these prediction tools in the EMR can provide significant cost savings and enhance patient care by reducing case cancellations.

# **REFERENCES**

1. Rozario N, Rozario D. Can machine learning optimize the efficiency of the operating room in the era of COVID-19? *Canadian Journal of Surgery*. 2020; 63(6):E527–E529.
2. Bartek MA, Saxena RC, Solomon S, et al. Improving Operating Room Efficiency: Machine Learning Approach to Predict Case-Time Duration. *Journal of the American College of Surgeons*. 2019; 229(4):346–354e3.
3. Huang CC, Lai J, Cho DY, Yu J. *A Machine Learning Study to Improve Surgical Case Duration Prediction*. June 12, 2020.
4. Martinez O, Martinez C, Parra CA, Rugeles S, Suarez DR. Machine learning for surgical time prediction. *Computer Methods and Programs in Biomedicine*. 2021; 208():106220.
5. Zhao B, Waterman RS, Urman RD, Gabriel RA. A Machine Learning Approach to Predicting Case Duration for Robot-Assisted Surgery. *Journal of Medical Systems*. 2019; 43(2):32.
6. Jiao Y, Sharma A, Ben abdallah A, Maddox TM, Kannampallil T. Probabilistic forecasting of surgical case duration using machine learning: model development and validation. *Journal of the American Medical Informatics Association*. 2020; 27(12):1885–1893.
7. 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: A Randomized Clinical Trial. *JAMA Surg.* 2021;156(4):315–321.
8. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024():e078378
9. Alotaibi FA, Aljuaid MM. A Comparison of Surgeon Estimated Times and Actual Operative Times in Pediatric Dental Rehabilitation under General Anesthesia. A Retrospective Study. J Clin Med. 2023 Jul 5;12(13):4493.
10. Smith, A. J., et al. (2021). Predicting surgical case durations using machine learning: A systematic review. *Surgical Endoscopy*, 35(8), 4482-4491
11. Eijkemans MJ, Van Houdenhoven M, Nguyen T, Boersma E, Steyerberg EW, Kazemier G. (2010) Predicting the unpredictable: A new prediction model for operating room times using individual characteristics and the surgeon’s estimate. Anesthesiology, 112(1):41-49.
12. Schneider A,  Wilhelm D,  Schneider MF,  Schuster T,  Kriner M,  Leuxner C,  Can S,  Fiolka A,  Spanfellner B,  Sitou W, Feussner H. (2011) Laparoscopic cholecystectomy – a standardized routine laparoscopic procedure: is it possible to predict the duration of an operation? Journal of Healthcare Engineering. 2(2): 287-298.
13. Wang J, Cabrera J, Tsui KL, Guo H, Bakker M, Kostis JB. (2020) Clinical and nonclinical effects on operative duration: evidence from a database on thoracic surgery. Journal of Healthcare Engineering. ID 3582796, 8 pages.
14. Kayıs E,  Khaniyev TT, Suermondt J, Sylvester K. (2015) A robust estimation model for surgery durations with temporal, operational, and surgery team effects. Health Care Management Sciences. 18:222-233.
15. Zain Z, Almadhoun MKIK, Alsadoun L, Bokhari SFH. Leveraging Artificial Intelligence and Machine Learning to Optimize Enhanced Recovery After Surgery (ERAS) Protocols. Cureus. 2024 Mar 21;16(3)
16. Stuart Geman, Elie Bienenstock, René Doursat; Neural Networks and the Bias/Variance Dilemma. *Neural Comput* 1992; 4 (1): 1–58.
17. Amin A, Cardoso SA, Suyambu J, Abdus Saboor H, Cardoso RP, Husnain A, Isaac NV, Backing H, Mehmood D, Mehmood M, Maslamani ANJ. Future of Artificial Intelligence in Surgery: A Narrative Review. Cureus. 2024 Jan 4;16(1):e51631.
18. Kelly, C.J., Karthikesalingam, A., Suleyman, M. *et al.* Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* **17**, 195 (2019).
19. Gupta, H., et al. (2023). Machine learning in surgery: Challenges and opportunities. *Surgical Innovation*, 30(2), 123-131.
20. Jiao Y, Xue B, Lu C, Avidan MS, Kannampallil T. Continuous real-time prediction of surgical case duration using a modular artificial neural network. Br J Anaesth. 2022 May;128(5):829-837.
21. Stucky, C., et al (2024). Surgical control time estimation variability: Implications for medical systems and the future integration of AI and ML models, *Perioperative Care and Operating Room Management*, 37.
22. Lau, H., et al (2010). Retrospective analysis of surgery postponed or cancelled in the operating room. *Journal of Clinical Anesthesia*, 22(4):237-40.

# **7 TABLES & FIGURES**

**Figure 1: Cohort Selection Flowchart**



1. Selected procedures: appendectomy, cholecystectomy, colectomy, gastric bypass, abdominal hernia repair, ileostomy closure, liver resection, mastectomy, parathyroidectomy, thyroidectomy, Whipple procedure.
2. Selected procedures: hiatal hernia, diaphragmatic hernia, paraesophageal hernia, radical cholecystectomy, exploratory laparotomy.

**Table 1: Baseline Characteristics of Study Population**

|  |  |
| --- | --- |
| **Patient Characteristics** | **Cohort (N=17246)** |
| Age, mean ± SD, y | 56.85 ± 15.95 |
| Sex, no. (%) |  |
| Male | 8189 (47.48%) |
| Female | 9042 (52.53%) |
| BMI, mean ± SD, kg/h2 | 30.54 ± 8.98 |
| ASA physical status classification, no. (%) |  |
| Class I | 1574 (9.13%) |
| Class II | 6184 (35.86%) |
| Class III | 8275 (47.98%) |
| Class IV | 1210 (7.02%) |
| Class V | 3 (0.02%) |
| Most responsible diagnosis (ICD-10), no. (%) |  |
| Digestive | 9791 (56.77%) |
| Neoplastic (Benign) | 932 (5.40%) |
| Neoplastic (Malignant) | 3404 (19.74%) |
| Endocrine/Metabolic | 2056 (11.92%) |
| External Causes | 837 (4.85%) |
| Other | 226 (1.31%) |
| **Surgical Characteristics** |  |
| Case service, no (%) |  |
| General Surgery | 14374 (83.35%) |
| Other | 2873 (16.65%) |
| Surgical encounter type, no. (%) |  |
| One day stay | 9116 (52.86%) |
| Same day admission | 7898 (45.80%) |
| Inpatient | 232 (1.35%) |
| Scheduled procedure, no (%) |  |
| Abdominal hernia repair | 5849 (33.92%) |
| Cholecystectomy | 3449 (20.00%) |
| Thyroidectomy | 2067 (11.99%) |
| Parathyroidectomy | 796 (4.62%) |
| Appendectomy | 184 (1.07%) |
| Gastric bypass | 837 (4.85%) |
| Mastectomy | 320 (1.86%) |
| Liver resection | 819 (4.75%) |
| Surgical approach, no. (%) |  |
| Open | 10810 (62.68%) |
| Laparoscopic | 6409 (37.16%) |
| Robotic | 27 (0.16%) |
| Anesthetic type, no (%) |  |
| General anesthetic | 16473 (95.52%) |
| Regional anesthetic | 481 (2.79%) |
| Monitored anesthetic care | 292 (1.69%) |

Abbreviations: SD, standard deviation; BMI, body mass index; ASA, American Society of Anesthesiologists; ICD-10, International Statistical Classification of Diseases and Related Health Programs 10th Revision.

**Table 2: Evaluation of Model Performance**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Scheduled Duration** | **Neural Network** | **Linear Regression** | **Ridge Regression** | **Lasso Regression** | **Random Forest** | **GBM** | **XGBoost** |
| MSE | 2498.4 ± 202.0 | **2469.3** ± **224.1** | 3137.5 ± 299.1 | 3131.2 ± 305.1 | 3134.1 ± 300.9 | 2567.3 ± 194.3 | 583.5 ± 245.6 | 2507.0 ± 214.6 |
| RMSE (minutes) | 49.9 ± 2.0 | **49.7** ± **2.3** | 56.0 ± 2.7 | 55.9 ± 2.7 | 55.9 ± 2.7 | 50.6 ± 1.9 | 50.8 ± 2.4 | 50.0 ± 2.2 |
| MAE | 35.3 ± 1.0 | **31.8** ± **1.2** | 36.9 ± 1.3 | 36.9 ± 1.3 | 36.9 ± 1.3 | 36.8 ± 1.3 | 32.4 ± 1.3 | 32.5 ± 1.4 |
| MAPE | 0.34 ± 0.01 | **0.26** ± **0.01** | 0.31 ± 0.01 | 0.31 ± 0.01 | 0.31 ± 0.01 | **0.26** ± **0.01** | 0.27 ± 0.01 | 0.27 ± 0.01 |
| R2 | **0.78** ± **0.03** | **0.78** ± **0.01** | 0.72 ± 0.02 | 0.72 ± 0.02 | 0.72 ± 0.02 | 0.77 ± 0.02 | **0.78 ± 0.02** | **0.78 ± 0.02** |

Abbreviations: GBM, MSE, mean squared error; RMSE, root mean squared error; MAE, mean absolute error; MAPE, mean absolute percentage error; *R*2, coefficient of determination.

**Table 4: Residual Analysis**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Scheduled Duration** | **Neural Network** | **Linear Regression** | **Ridge Regression** | **Lasso Regression** | **Random Forest** | **GBM** | **XGBoost** |
| Mean residual, minutes ± SD | -18.52 ± 36.72 | -0.37 ± 40.05 | -1.72 ± 40.47 | -2.05 ± 40.20 | -2.13 ± 40.92 | -2.69 ± 44.95 | -2.67 ± 44.87 | -2.63 ± 44.94 |
| Median residual, minutes ± IQR | -19.00 ± 39.00 | -3.69 ± 37.20 | -4.52 ± 47.15 | -4.66 ± 47.18 | -4.83 ± 46.96 | -4.84 ± 37.82 | -5.62 ± 38.73 | -5.84 ± 38.59 |
| t-statistic | -52.956 | -0.960 | -6.155 | -6.294 | -6.242 | -4.455 | -5.456 | -5.346 |
| p-value | 0.00e+00\*\* | 3.37e-01 | 7.75e-10\*\* | 3.22e-10\*\* | 4.48e-10\*\* | 8.47e-06\*\* | 4.97e-08\*\* | 9.17e-08\*\* |

Abbreviations: GDM, gradient boost machine; SD, standard deviation; IQR, interquartile range Comparison of residuals for different models using t-test. Significance levels are indicated by ∗(*p <* 0*.*05) and ∗∗ (*p <* 0*.*01).

**Table 3: Comparison of Machine Learning Models and Scheduled Duration**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Comparison** | **t-statistic** | **p-value** | **Better Model** |
| MSE | Linear Regression | -5.60 | 4.18e-05 \*\* | Scheduled duration |
| Ridge Regression | -5.47 | 5.60e-05 \*\* | Scheduled duration |
| Lasso Regression | -5.55 | 4.69e-05 \*\* | Scheduled duration |
| Random Forest | -0.78 | 4.47e-01 | Equivalent |
| GBM | -0.83 | 4.15e-01 | Equivalent |
| XGBoost | -0.09 | 9.32e-01 | Equivalent |
| **Neural Network** | **0.32** | **7.56e-01** | **Equivalent** |
| RMSE | Linear Regression | -5.69 | 2.82e-05 \*\* | Scheduled duration |
| Ridge Regression | -5.56 | 3.81e-05 \*\* | Scheduled duration |
| Lasso Regression | -5.64 | 3.17e-05 \*\* | Scheduled duration |
| Random Forest | -0.78 | 4.43e-01 | Equivalent |
| GBM | -0.83 | 4.15e-01 | Equivalent |
| XGBoost | -0.09 | 9.32e-01 | Equivalent |
| **Neural Network** | **0.32** | **7.56e-01** | **Equivalent** |
| MAE | Linear Regression | -2.96 | 8.87e-03 \*\* | Scheduled duration |
| Ridge Regression | -3.09 | 6.67e-03 \*\* | Scheduled duration |
| Lasso Regression | -2.88 | 1.04e-02 \* | Scheduled duration |
| Random Forest | 5.68 | 2.56e-05 \*\* | Random Forest |
| GBM | 4.19 | 7.42e-04 \*\* | GBM |
| XGBoost | 5.31 | 6.28e-05 \*\* | XGBoost |
| **Neural Network** | **7.01** | **1.81e-06 \*\*** | **Neural Network** |
| MAPE | Linear Regression | 7.52 | 7.79e-07 \*\* | Linear Regression |
| Ridge Regression | 7.16 | 1.35e-06 \*\* | Ridge Regression |
| Lasso Regression | 7.68 | 6.23e-07 \*\* | Lasso Regression |
| Random Forest | 20.34 | 3.84e-12 \*\* | Random Forest |
| GBM | 14.69 | 1.92e-11 \*\* | GBM |
| XGBoost | 18.70 | 4.79e-12 \*\* | XGBoost |
| **Neural Network** | **20.95** | **3.53e-13 \*\*** | **Neural Network** |
| R2 | Linear Regression | 5.49 | 4.35e-05 \*\* | Scheduled Duration |
| Ridge Regression | 5.43 | 4.91e-05 \*\* | Scheduled Duration |
| Lasso Regression | 5.48 | 4.55e-05 \*\* | Scheduled Duration |
| Random Forest | 0.55 | 5.88e-01 | Equivalent |
| GBM | 0.66 | 5.19e-01 | Equivalent |
| XGBoost | -0.03 | 9.78e-01 | Equivalent |
| **Neural Network** | **-0.42** | **6.81e-01** | **Equivalent** |

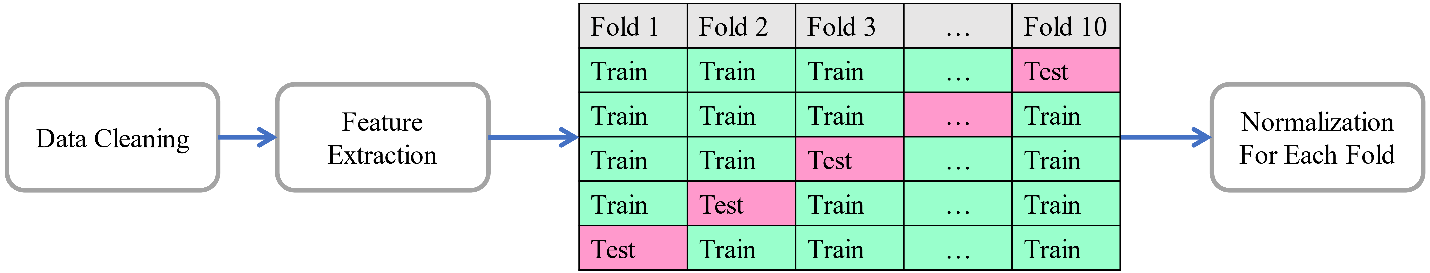
Abbreviations: GBM, MSE, mean squared error; RMSE, root mean squared error; MAE, mean absolute error; MAPE, mean absolute percentage error; *R*2, coefficient of determination.

**8 SUPPLEMENTARY MATERIAL**

**Supplementary Table 1: Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Cohort (N=17,246)** |
| age\_at\_discharge | Numerical | Age of patient at time of discharge. Excludes <18, >130 years. |
| sex | Categorical | Genotypical sex of patient. Transgender, intersex collapsed into 'Other'. |
| average\_BMI | Numerical | Average BMI (kg/m²) during encounter. Excludes <5, >200; missing data imputed with mean. |
| asa\_score | Ordinal | American Society of Anesthesiologists (ASA) co-morbidity score. |
| surgical\_encounter\_type | Categorical | Type of surgical encounter. |
| case\_service | Categorical | Main surgical service. |
| scheduled\_procedure | Categorical | Name of scheduled procedure. Selects general surgery procedures. |
| procedure\_approach | Categorical | Approach for the scheduled procedure (e.g., Open, Laparoscopic). |
| diagnosis\_category | Categorical | ICD10 category of the most responsible diagnosis. Based on most responsible diagnosis. |
| anesthetic\_type | Categorical | Type of anesthesia used. Collapsed into General, Regional, Sedation, Local. |
| day\_of\_year | Numerical | Day of the year on which the case is scheduled. Based on case date. |
| day\_of\_week | Ordinal | Day of the week on which the case is scheduled. Based on case date. |
| weekend\_indicator | Binary | Indicates if the case is scheduled on a weekend. Based on case date. |
| surgical\_location\_hospital | Categorical | Hospital where the surgery takes place. Based on surgical location. |
| surgical\_location\_service | Binary | Surgical specialty of the OR. Based on surgical location. |
| primary\_surgeon\_id | Categorical | Unique identifier for the primary surgeon. Categories with <1% frequency collapsed into 'Other'. |
| scheduled\_surgeon | Binary | Indicates if there is an additional scheduled surgeon (as surgical assist). |
| first\_case\_of\_day | Binary | Indicates if the case is the first scheduled of the day in the assigned OR. |
| last\_case\_of\_day | Binary | Indicates if the case is the last scheduled of the day in the assigned OR. |
| morning\_scheduled | Binary | Indicates if the case is scheduled to start in the morning (AM). Based on scheduled start time. |
| actual\_case\_time\_minutes | Numerical | Duration of the case in minutes. Excludes >8000 minutes. |

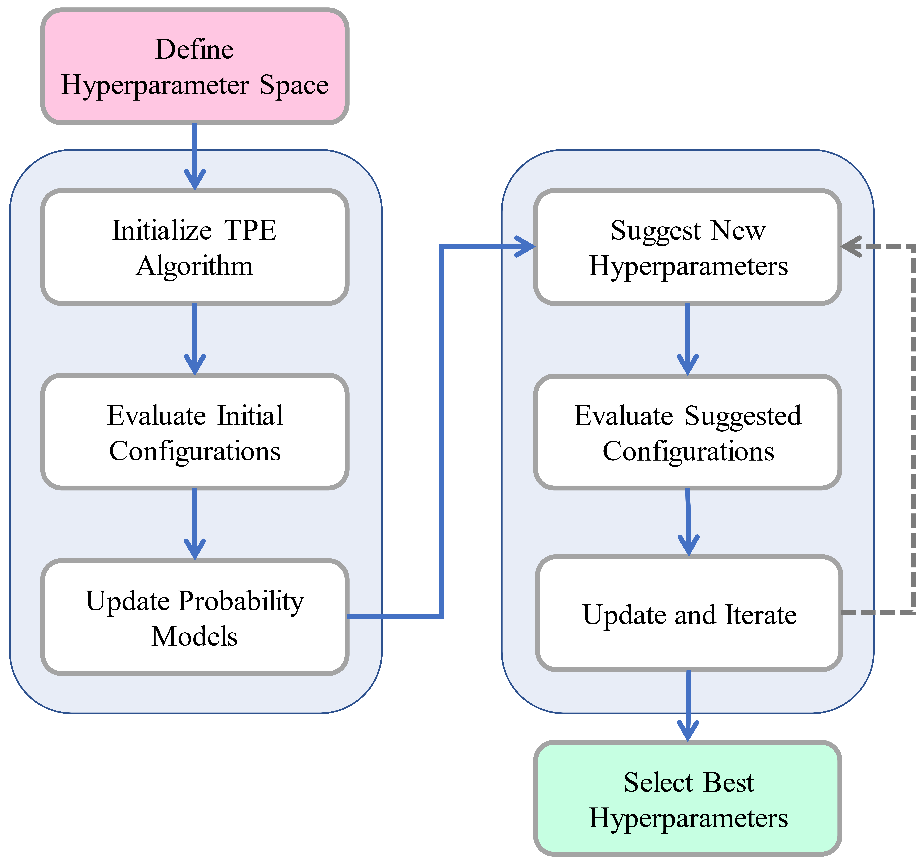
**Supplementary Figure 1: Data preparation & pre-processing workflow**



**Supplementary Figure 1**: Workflow of data preparation and pre-processing, including data cleaning, feature extraction, normalization, and 10-fold cross-validation.

Supplementary Figure 3: Hyperparameter optimization workflow

**Supplementary Figure 2: Hyperparameter Optimization**



**Supplementary Figure 3:** Flowchart of the hyperparameter optimization process using the Tree-structured Parzen Estimator (TPE) algorithm, from defining the hyperparameter space to selecting the best configuration.

**Supplementary Table 2: Hyperparameter Ranges for Optimization**

|  |  |  |
| --- | --- | --- |
| **Model Type** | **Hyperparameter** | **Optimization Range** |
| Linear Regression | none | N/A |
| Ridge Regression | alpha | 1e-5 to 1e2 (log scale) |
| Lasso Regression | alpha | 1e-5 to 1e2 (log scale) |
| SVR | C | 1e-3 to 1e2 (log scale) |
| epsilon | 1e-3 to 1.0 (log scale) |
| kernel | {'linear', 'poly', 'rbf'} |
| Random Forest | n\_estimators | 100 to 1000 |
| max\_depth | 3 to 30 |
| min\_samples\_split | 2 to 10 |
| min\_samples\_leaf | 1 to 4 |
| GBM | n\_estimators | 100 to 1000 |
| learning\_rate | 1e-3 to 1.0 (log scale) |
| max\_depth | 3 to 30 |
| subsample | 0.5 to 1.0 |
| XGBoost | n\_estimators | 100 to 1000 |
| learning\_rate | 1e-3 to 1.0 (log scale) |
| max\_depth | 3 to 30 |
| subsample | 0.5 to 1.0 |
| colsample\_bytree | 0.5 to 1.0 |
| Neural Network | units\_l1 | 64 to 512 |
| activation\_l1 | {'relu', 'leaky\_relu', 'elu'} |
| dropout\_l1 | 0.0 to 0.5 |
| l2\_l1 | 1e-6 to 1e-2 |
| n\_layers | 2 to 5 |
| optimizer | {'adam', 'nadam', 'rmsprop'} |
| learning\_rate | 1e-5 to 1e-2 (log scale) |
| batch\_size | 16 to 128 |

**Supplementary Material 1: Model Evaluation Metrics**

**Mean Squared Error (MSE)** measures the average of the squares of the errors between the actual and predicted surgery times. MSE is particularly sensitive to larger errors, making it useful for identifying models that minimize significant prediction discrepancies.

where represents the actual surgery time, is the predicted surgery time, and *n* is the number of observations.

**Root Mean Squared Error (RMSE):** This metric is the square root of the MSE, providing an error measure in the same units as the case time of surgery (minutes). RMSE is useful for understanding the typical magnitude of errors in the predictions.

**Mean Absolute Error (MAE):** This metric measures the average magnitude of the errors in predicting surgery time, without considering their direction. It is less sensitive to outliers compared to MSE and RMSE, making it a robust measure of model accuracy.

**Mean Absolute Percentage Error (MAPE):** This metric expresses the prediction accuracy as a percentage by measuring the average of the absolute percentage errors between the actual and predicted surgery times. It provides a relative measure of error, which is useful for comparing the accuracy across different scales.

**Coefficient of Determination (***R*2**):** This metric indicates the proportion of variance in the actual surgery times that can be explained by the predictive model. An *R*2 value closer to 1 suggests that the model accurately captures the variability in surgery times.

where is the mean of the actual surgery times.

1. Department of Anesthesiology and Pain Medicine, University of Alberta [↑](#footnote-ref-1)
2. Department of Medicine, Western University [↑](#footnote-ref-2)
3. Department of Electrical and Computer Engineering, Western University [↑](#footnote-ref-3)
4. Department of Surgery, Western University [↑](#footnote-ref-4)
5. \* Corresponding author: Dr. Ahmad Elnahas, MD, FRCSC

   London Health Sciences Center, Western University

   339 Windermere Rd., London ON, N6A 5A5

   (t) 519.663.3612, (e) [Ahmad.Elnahas@lhsc.on.ca](mailto:Ahmad.Elnahas@lhsc.on.ca) [↑](#footnote-ref-5)