Predicting orthopedic surgery times using machine learning

Senobar Kheradmand Igder  
Persian Gulf University

Khodakaram Salimifard (salimifard@pgu.ac.ir)  
Persian Gulf University

Seyed Mohammad Tahami  
Shiraz University of Medical Sciences

Reza Mohammadi  
University of Amsterdam

Alex Kuiper  
University of Amsterdam

Shahin Heidari  
Shiraz University of Medical Sciences

Research Article

Keywords: surgery time, orthopedic, prediction, machine learning, operating room

Posted Date: January 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2409909/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

The operating rooms within the surgical unit take center stage in a hospital. The fact that, in practice, actual durations of surgery do not coincide with their allotted times yields extra costs; for example, earliness results in unutilized operating room time, and lateness incurs extra waiting for patients. Various machine learning methods are employed to predict surgery times in a hospital. The data used stems from the Shahid Chamran Trauma educational-medical hospital (Shiraz, Iran) from 2018 until 2021. The performances across the four methods, linear regression, recursive partitioning, support vector machine, and XGBoost, are compared using established accuracy and relevant healthcare operational metrics. The predicted surgery times vary per algorithm, but the differences are minor. Among the methods, linear regression shows the best performance. Linear regression, which also provides explanatory insights, outperforms the other approaches for predicting surgery times. Furthermore, the study shows that using machine learning models is a promising avenue to improve the prediction of operation time and generate more efficient and effective operating room schedules.

1. Introduction And Related Works

Healthcare services are under increasing pressure to improve their operations. On the one hand, hospitals aim to reduce costs and improve their financial situation and outlook. On the other hand, they try to increase the level of satisfaction for patients. Operating rooms are a hospital's highest cost and revenue; therefore, improving the operative processes significantly impacts the hospital's performance [1]. While there is a great variety in healthcare systems, governmental agencies across the globe are stimulating to improve healthcare operations management to reduce costs, see [2]. But also one of the latest OECD reports [3] states: "Reforms to improve economic efficiency are critical".

Although the focus is often placed on efficiency, it is reported that 10–20% of all scheduled surgeries are canceled or rescheduled, mainly because of hospital-related reasons: over-runs of previous surgeries, procedural grounds (e.g., patient not ready, no surgeon) and no postoperative beds available [4]. Operating rooms play a critical role in a hospital. There are only a limited number of highly cost-intensive operating rooms available for which there is dedicated supporting staff and specialist equipment [5][5](Zhu et al., 2019)[5] but also due to the ramifications of inaccurate OR scheduling, see [6]. For scheduling, the available time is often divided among different medical specialties (e.g., orthopedics) in which patients are scheduled and sequenced [7].

Jebali et al. [8] also outline that operating room scheduling is a complex and multifaceted problem. Ultimately, it results in specific timings when patients undergo surgery. To build up a schedule, one should have estimates of the surgery durations. Naturally, one would want these estimations as accurate as possible, having low variability, but extreme deviations should also be rare [9]. At the same time, it is widely known that the duration of surgery is affected by many different factors, such as the type of surgery, experiences of the surgical team, and the patient's pathology, which therefore make up a large part of the variability [10]. In line with this pursuit, according to a review [11], an operating room schedule that incorporates information that can be retrieved from operating room information systems improves the efficiency of the operating room and related resources.

Surgery schedules are mainly built based on the surgeon's time predictions. These predictions are often based on overall averages [12] or moving averages (over 5 to 10 previous procedures) [13], which can be manually adjusted to incorporate specific clinical conditions. If the surgery duration is less than the actual time in practice, it might lead to idle time for resources and equipment. On the other hand, if the estimated surgery duration is longer than realized, it potentially prolongs patients' waiting times. Also, it can accumulate extra, unplanned costs such as session overruns, see [6] and [14]. This mismatch can be disentangled into variations that can be incorporated into a prediction model, such as patient characteristics, type of operation, surgeon, and randomness that cannot be attributed to specific factors, which, therefore, cannot be part of the prediction model.

To demonstrate the potential of prediction in an orthopedic cluster, we consider the work by [15] in which nine fellowship-trained orthopedic surgeons at a single institution were asked to estimate their operative and total room times over three months. The study aimed to uncover the planning fallacy (the tendency to underestimate actual time durations) in orthopedic surgery. However, focusing on the operative times, the data, consisting of 759 cases, did not support this fallacy. It was also shown that surgeons could improve their time estimations even further by allowing them to learn from the differences between their predicted and actual durations. In another, but similar [16], it was found that surgeons accurately predicted the duration of 20 hysterectomies (26.7%), whereas the system obtained a score of only 9.3%. So, there is critical information to be learned. Still, instead of asking specialists, we aim to develop a machine learning algorithm that is objective and uniformly outperforms specialists in terms of accuracy.

Various research works have focused on predicting surgery times in the operating room. To name a few, [17] showed that it is advantageous to leverage information such as patient age, gender, morbidity, anesthesiologist identity, and surgery location to improve time estimations. For his study, he compared the accuracy of four standard machine learning: linear regression, nearest neighbors, regression trees, and support vector regression, reporting that the inclusion of the additional variables led to a 20% accuracy increase. In another effort to improve surgery schedules, [18] used a combination of a data mining model and an optimization algorithm. They applied three different data mining algorithms to predict the duration of surgeries and compared them with the estimates made by surgeons. Also, the table in [13] shows that machine learning can outperform the predictions that are provided by expert schedulers for nearly all specialties. After some preprocessing steps, two linear regression models are employed (a simple model and one with interactions) to predict the surgery times across fifteen specialties. Notable is the vast number of data in this research, using variables extracted from a data mining processing of the hospital's clinical histories (around 63,000 surgeries over 39 months are in the dataset, with six variables: surgery type, procedure type, physical status, patient age, surgery scope, and specialty).

Martinez et al. [19] used linear regression, support vector machines, regression trees, and bagged trees to predict surgical time durations at a tertiary referral university hospital in Bogotá, Colombia. The data consisted of primary patient and surgeons' information (patient’s age and destination, and surgeon’s ID
and experience). The algorithms mentioned above resulted in varying prediction performances, ranging from 26 to 37 root mean squared errors but bagged trees, which resulted in a 40% improvement in the prediction scores, yielding the best score.

This research aims to employ machine learning algorithms on an actual dataset to show the gain that can be obtained by using such a method. In addition, it can serve as a starting point for rendering better surgery schedules. In the next section, we define this surgery scheduling problem. Section 3 elaborates on the dataset, where it is obtained, and how it is prepared. In Section 4, we introduce the algorithms to be applied and the performance metrics on which our assessment of the machine learning models will be based. We present our findings in Section 5, after which, in Section 6, we discuss and conclude.

2. The Surgical Time Prediction Problem

Predicting surgical duration is the basis of resource management and operating room planning in the surgical department, and inaccurate scheduling leads to extra costs [6]. The outcomes of the surgery scheduling process are the sequence and timings of patients to the operating rooms. While trying to solve this problem, many requirements should be simultaneously considered, i.e., resource constraints, ward availability, and intensive care occupations [8]. Still, prediction models can play an essential role in the decision-making process around surgery’s start and end times as one has better estimates about the actual time durations needed to build up a schedule.

The benefits of having more accurate surgery timings are foremost a reduction of earliness and lateness of surgeries. Moreover, more precise estimates can lead to more effective OR room use. Knowing how long procedures will take helps to determine how many surgeries can be done in a single day. Therefore, it also improves the loading of operating rooms, minimizing unplanned underutilization of operating room time, and at the same time, it can reduce unanticipated session overruns. To this end, this study proposes a data-driven algorithm that predicts the duration of elective orthopedic surgeries more accurately than current estimates.

3. Dataset

The data for this case study originates from the Shahid Chamran Orthopedic Hospital in Shiraz, Iran. The hospital uses different experimental methods to estimate surgery times and relies on the insights to draft operating room schedules. The data is extracted from the hospital’s Electronic Health Record system (EHR) database and operating room records. The data cover all surgeries from 2018 to 2021.

The EHR captures information about the patient and the surgery performed, the patient (gender, age), the surgery (surgery code, surgical duration, anesthesia time, k surgery, k anesthesia, and the target variable surgery time), and the operative team (surgeon). Variables were selected using a stepwise regression method. Stepwise regression is used when there is uncertainty in deciding which predictor variables to be included in the regression model. The dataset consists of 31,232 records, of which the majority are categorical. This is a challenge in developing machine learning algorithms to predict surgery times requiring various encoding methods to solve it. Our plans and procedures will be explained in the following section.

Hospitals within the United States commonly use the current procedural terminology (CPT) coding system developed by American Medical Association (AMA) to identify specific procedures performed during surgery [20]. In 1991, the AMA began reimbursing hospitals based on the relative value unit. A standardized index associated with each CPT captures the staff workload for a given procedure [21]. The calculated RVU is declared as a fee called the k index. In this research, k_surgery and k_anesthesia are the two variables defined using the RYU. Table 1 shows variables selected after a stepwise regression was applied.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Data type</th>
<th>No. of categories</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Categorical</td>
<td>2</td>
<td>Patient sex</td>
</tr>
<tr>
<td>Surgery class</td>
<td>Categorical</td>
<td>14</td>
<td>Grouped by surgery</td>
</tr>
<tr>
<td>Surgery code</td>
<td>Categorical</td>
<td>13</td>
<td>Surgery code</td>
</tr>
<tr>
<td>K surgery</td>
<td>Categorical</td>
<td>10</td>
<td>K surgery</td>
</tr>
<tr>
<td>K anesthesia</td>
<td>Categorical</td>
<td>10</td>
<td>K anesthesia</td>
</tr>
<tr>
<td>Surgeon</td>
<td>Categorical</td>
<td>14</td>
<td>Surgeon name</td>
</tr>
<tr>
<td>Surgery time</td>
<td>Numerical</td>
<td>-</td>
<td>Surgery time</td>
</tr>
<tr>
<td>Age</td>
<td>Numerical</td>
<td>-</td>
<td>Patient age</td>
</tr>
</tbody>
</table>

3.1. Data cleaning

To work with the data, several preprocessing steps must be undertaken. Figure 1 illustrates the first data-cleaning steps. The outlined procedure consists of four stages: selection, filtering, merging, and finishing. The dataset was extracted from the hospital database and converted into an Excel format. After
applying different filters, the number of records was 16,450. The records with unusual surgery times were removed, and the number of records at the end of
this filtering was reduced to 16,224 surgeries.

3.2. Variable encoding

All variables in this research are categorical except age and time of anesthesia. Most ML algorithms, such as neural networks, use numerical variables to
predict a target variable. Therefore, categorical variables should be converted into numerical variables to be used as input variables in the model. One Hot
Encoding (OHE) is a useful tool in feature engineering for machine learning models. This method creates a design (or model) matrix by expanding factors
to a set of dummy variables (depending on the contrasts) and expanding interactions similarly. Variables including gender, surgeon, surgery_code,
data_class, anesthesia_time, k_surgery, and k_anesthesia were coded using OHE.

3.3. Feature selection

As the literature describes, features are the attributes or properties of the observations required for machine learning. Finding the key features takes place in
the feature selection. Here, the feature selection was performed by using stepwise regression. The selected variables are gender, surgeon, surgery_code,
data_class, anesthesia_time, k_surgery, k_anesthesia, and age. Variables k_surgery and k_anesthesia were grouped using the method suggested in the CPT
[22]. The interested reader is referred to [23] for an explanation of the model of such a selection procedure.

The surgery duration is considered to start at the beginning of anesthesia and last until the patient is transferred to the recovery room. Variables were
selected with the advice of the hospital orthopedic lead surgeon. Then, patients’ records were splitted into several groups based on the part of the body on
which the operation was done. To select features, all variables were entered in a stepwise regression model, and accordingly, variables with a significance
of < 0.01 were selected for the prediction model.

4. Machine Learning Algorithms

This research used four machine learning algorithms to predict surgery time. Specifically, algorithms include RPART, regression tree, support vector
machine (SVM), and XGBoost. These algorithms are known as supervised machine learning methods [24]. To apply those four algorithms, a set of steps
must be taken. Figure 2 provides an overview of the framework in which we trained and tested the machine learning models.

4.1. Linear Regression

Linear regression, for example [25], attempts to model the relationship between two or more variables by fitting a linear equation to the observed data. The
variable of interest is considered the dependent variable, while the others in such a model are considered independent variables that affect the dependent
variable.

4.2. RPART

The RPART (Recursive PARTitioning) algorithm [26] builds a classification or regression model. The RPART algorithm works by splitting the dataset
recursively. At each split the dataset is separated into two, which continues until a predetermined termination criterion is reached. Hence, a resulting model
can be represented as a binary tree.

4.3. Support Vector Regression Machines

Support Vector (Regression) Machines, often abbreviated to SV(R)M, e.g., [27] and [28], both use such a supervised machine learning algorithm to predict
discrete (numerical) values. For classification, these models can be understood as finding a line (or hyperplane in higher dimensions) that separates the
data; for regression, the same idea can be applied after introducing a concept of measuring prediction error, e.g., through a loss function.

4.4. Extreme Gradient Boosting

Developed by Chen and Guestrin [29], Extreme Gradient Boosting (known as XGBoost) is a variant of gradient boosting and a supervised learning
algorithm. Data scientists widely use XGBoost as it often achieves the best performance in many classification and regression problems. XGBoost provides
a parallel tree-boosting algorithm and thus uses an ensemble of estimates to finally output a prediction.

4.5. Training methodology

We split the dataset into a train and test according to the 80 - 20 rule: 80% of the data are selected as the train set, used to fit and fine tune the machine
learning model. The remaining data (20%) as the test set. The test set is used to evaluate the machine learning model on unseen data and allows a fair
comparison across several machine learning models. Mean Squared Error (MSE) is a frequently used loss function for regression problems. The loss is
defined as the mean of the squared differences between actual and predicted values

\[ MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i' - y_i)^2, \]

1
where \( N \) is the sample size, and, for each observation \( i \), \( y_i' \) is the predicted value of the surgery time and \( y_i \) its actual value. The Root Mean Squared Error (RMSE) is a slight variant of the above but has the advantage that it has the same unit of measurement as the outcome variable. It takes the square root of the MSE and is thus calculated as

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i' - y_i)^2}{N}}.
\]

We train and test the machine learning models under two different approaches. The first approach is to apply the prediction model to the data altogether. However, at the request of hospital authorities (e.g., senior orthopedic surgeons), we segment the data such that a unique model is fitted for each surgery type.

**Approach 1 (All surgeries)**

We used all data to predict surgery time without segmenting the dataset into different surgery types. The dataset consists of the existing data shown in Table 1. All surgery types, all surgeons, surgery code, \( k \) surgery, and anesthesia time are used in the prediction model.

**Approach 2 (Segmented surgeries)**

The dataset is split into nine segments; each consists of data belonging to one specific orthopedic surgery category. Then, we trained and tested the ML models on each of these nine categories. The splits are based on the type of surgery on a specific section of the body’s skeleton. The nine orthopedic surgery types are spine, femur, hand, forearm, foot, knee, shoulder, leg, and other surgeries. The surgery types and the dataset for each prediction model are shown in Table 2. It also shows the number of different levels for the four categorical variables under the segmentation and the number of observations.

The machine learning algorithms were implemented using RStudio software installed on a windows 10 operating system. The hardware platform had 64 GB RAM, CPU Intel Core i7-1165G7 up to 4.7GHz, 1TB disk size, and an Intel Iris Xe Graphics video card.

### 5. Results And Findings

This section presents the results of developing prediction models using four machine learning algorithms. Using the performance measures already introduced, we aim to identify the algorithm that has the best performance. Following that, we compared the results of the machine learning algorithm against the actual surgery time. Table 3 shows that approach 1 (All surgery) consists of all 16224 records in the dataset, while approach 2 (segmented surgeries) consists of nine different sets of records extracted from the dataset.

The table below shows the resulting RMSE for the prediction models.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Approach 1</th>
<th>Approach 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>27.12367**</td>
<td>19.17466*</td>
</tr>
<tr>
<td>RPART</td>
<td>29.97946</td>
<td>21.91155**</td>
</tr>
<tr>
<td>SVRM</td>
<td>81.94458</td>
<td>58.76747</td>
</tr>
<tr>
<td>XGBoost</td>
<td>26.86978**</td>
<td>31.73483</td>
</tr>
</tbody>
</table>

Table 3 shows the RMSE results of the four machine-learning algorithms for ten prediction models. The symbols “*” and “**” in each column indicate the best and second-best algorithms in terms of performance. Among the four algorithms, linear regression (LR) performs remarkably well. The only surgery
type for which the LR is placed second, outperformed by XGBoost, is in the "All" category, i.e., under Approach 1, in which the surgery types are jointly considered in a single prediction model. So, in terms of RMSE scores, RPART and LR generally obtain the best performance, and, considering the prediction errors, they are at most 41 minutes off.

<table>
<thead>
<tr>
<th>Performance measures</th>
<th>Approach 1</th>
<th>Approach 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-squared</td>
<td>0.8501</td>
<td>0.8302</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.8494</td>
<td>0.8212</td>
</tr>
<tr>
<td>Residual standard error</td>
<td>24.45</td>
<td>20.04</td>
</tr>
</tbody>
</table>

Table 4 shows the results of R-squared, adjusted R-squared, and residual standard error. There is not much difference between R-squared and adjusted R-squared. In Approach 1, the value of R-squared is 0.8501, and the residual standard error is 24.45. In Approach 2, shoulder and spine categories have the maximum value for R-squared. The value of R-squared for shoulder surgery is 0.888 and for spine surgery is 0.8505. The residual standard errors for the two surgeries are 19.35 and 42.15, respectively. According to Table 3 and Table 4, although the spine category has a high R-squared value, it does not have a low RMSE. It is due to the fact that this category there is a because of the high residual standard error. Also, the shoulder category stands out with a high R-squared value, which can be explained by the fact that many different procedures can be reasonably well predicted.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Approach 1</th>
<th>Approach 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Mean</td>
<td>106.9</td>
<td>127.8</td>
</tr>
<tr>
<td>LR Mean</td>
<td>107.2</td>
<td>131.3</td>
</tr>
<tr>
<td>Actual Worst</td>
<td>495.0</td>
<td>495.0</td>
</tr>
<tr>
<td>LR Worst</td>
<td>424.6</td>
<td>307.5</td>
</tr>
</tbody>
</table>

In Table 5, we make an operational judgment on the performance of the prediction model. Since the LR produces the best performance, we have selected the LR as the prediction algorithm for surgery time. Under Approach 1, all surgeries with all features have been used to create the prediction model.

Figure 3 shows the surgical times predicted by linear regression against actual surgical time. The LR model is perfectly placed within the observations making the RMSE small. Note that the observed values are rounded to 5 minutes, as going along the diagonal in the figure, the dots seem clustered horizontally. Furthermore, notice the isolated cases quite far off, approximately 10 out of the total 3,244 of the test set. Applying Approach 2, surgeries are classified into nine surgery types. Accordingly, nine prediction models are shown in Fig. 4, which do not display any irregular patterns. Comparing the difference between the means of the predictions and the actuals in Table 5, we see that they never exceed 5 minutes.

In addition, considering the 'worst-case' scenario where the surgery takes the most time, we find that the model can detect these extremely lengthy surgeries. As seen in Table 5, the maximum actual worst-case is 495 minutes for Approach 1, while the worst-case predictions produced by LR algorithm Approach 1, knee and spine surgeries, are 424.6, 307.5, and 446 minutes, respectively. Overall, we conclude that the models are adequate to predict surgery times of orthopedic cases, and one can opt to put all data together or to segment per surgery type.

6. Conclusion And Discussion

Operating rooms play a vital and critical role in hospitals as these resources are extremely expensive [1]. Surgery scheduling is an important planning tool to ensure the good use of these scarce resources. Not only to be able to offer the services dependably without being too early or late but also to ensure that operating rooms are used efficiently. Eijkemans et al. [30] report that reliable predictions are an important prerequisite in this pursuit. The research proposed using machine learning algorithms to predict surgery times in orthopedic hospitals. In detail, we compared linear regression, RPART (Recursive PARTitioning), SVM (support vector machines), and XGBoost. We establish that linear regression yields the best prediction scores in the comparison. Besides its excellent performance, regression’s straightforward interpretability makes it attractive to be deployed in practice. In addition, we note that the prediction error is low across the different types of surgeries, which underpins that the model is not biased. In detail, for example, for shoulder surgery, the unexplained variation is only 12%.

A limitation of this study was that the data contained a lot of missing values and outliers. To properly use machine learning algorithms that predict the results more carefully requires quality data 25% of data was removed. Considering the dataset in more detail, while preprocessing the dataset missing values and outliers have resulted in a reduction of 25% of the original data, which might lead to biased results. Furthermore, due to the COVID-19 outbreak falling in the study period, the hospital was treating patients infected by COVID-19, which made the number of surgeries in some specific surgery types
relatively low. We also remark that the dataset focuses on an orthopedic hospital in Iran. Although orthopedic treatments are quite standardized, differences across countries can occur. We, however, believe that the same performance can be obtained for other specialties.

In predicting orthopedic surgery durations, we considered the surgery duration from the onset of anesthesia until the surgery was finished. Therefore, a natural extension is also to include and predict other important timed activities, e.g., recovery time, operating room cleaning, and patient preparation. Of course, these external activities do not directly affect the operating room’s efficiency but can, of course, affect a session indirectly. In addition, using machine learning models in surgery scheduling problems in literature [1] is another logical avenue for further research. The revolving decisions in the surgery scheduling problem will then be made based on more precise estimates such that it likely results in less costs (less waiting and idleness). Furthermore, with more accurate predictions, one can also reconsider which patients to schedule first, i.e., the sequencing decision, see also [8]. So, the proposed method is a tool to generate more efficient and patient-friendly schedules.

**Declarations**

**Availability of data and material**

The datasets generated during and analyzed during the current study are not publicly available due to a non-disclosure agreement.

The R code of this research can be accessed at https://github.com/sekheradmand/prediction-time.

**Authors’ contributions**

S.K.I., K. S., R. M. made the conceptualization; S. H., S. M. T., S. K. I. prepared data; S. K. I., R. M., K. S., and A. K. did the methodology; S. K. I., K. S., and R. M. did the formal analysis and investigation; S. K. I., K. S., A. K., R. M., and S. M. T. wrote the main manuscript; and the research was supervised by K. S.

**Ethics Statement**

The studies involving human participants were reviewed and approved by the Ethics Committee of the Bone and Skeleton Research Center of the Shiraz University of Medical Science.

**Funding**

This study has received no funding from any third party.

**Competing Interests**

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

**References**


Data cleaning procedure.
Figure 2
Surgical Time Prediction Framework.

Figure 3
Predicted versus actual surgery time using the LR algorithm for Approach 1.
Figure 4

Plots of the predicted versus the actual surgery times for Approach 2: segmented surgeries.