# Construction of a Multiple Linear Regression Model for the Prediction of Surgical Times in the Orthopaedic Speciality of Hospital da Luz of Lisbon

Rita Cardoso Lourenço Silva Fernandes

Master in
Industrial Engineering and Management

## Abstract

The operating room (OR) is among the highest hospital revenue generators while also accounting for an equally high cost of use. Thus, optimising the use of the OR becomes vital for the provision of efficient and cost-effective healthcare. A fundamental step towards achieving optimal scheduling of surgical procedures is to obtain accurate estimates of their duration.
This master's thesis focuses on the study and construction of a resolution approach that can be applied as a tool to provide more accurate estimates on the use of the OR than the predictions currently made by the surgeons of the orthopaedic specialty at Hospital da Luz of Lisbon.
Based on a literature review, linear regression was selected as the method to be applied to the problem in question. Three different scenarios were built (aggregate model, model per procedure and model per surgeon) to understand which of the three provides forecasts that are closer to what is observed.
All models could outperform the predictions estimated by the surgeons, representing a preferable alternative to the currently used method. However, it was possible to confirm the best performance of the aggregate model compared to the others.

**Keywords:** Operating Room, Efficiency, Surgery Case Duration, Multiple Linear Regression

## 1. Introduction

The operating room (OR) is a critical resource that represents more than 40% of the total revenue of a hospital and an equally high proportion of its total expenditure, which makes it the most expensive unit but also the one with the highest source of income (Denton *et al.*, 2007). Thus, efficient management of this unit is fundamental when hospitals or other health services aim to maximise their results with the existing resources.
This resource is sometimes unpredictable and has multiple factors that may interfere with its efficiency and make complex and challenging the exercise of sequencing and scheduling of surgeries correctly to ensure efficient use of resources and to avoid under or overuse of the OR, which also can have an undesired impact on patient waiting times. One of the most important components in this exercise is the duration of each surgery, which is difficult to predict (Lee *et al.*, 2019). Therefore, when asked to present the expected time for each procedure, surgeons may overestimate or underestimate the duration of a surgery. When surgeries take longer than expected, the procedures that follow may be delayed (resulting in reduced starts at scheduled time with additional demands on workers' hours) or cancelled (creating throughput issues and undermining the value of the customer's hospital experience). On the other hand, overly conservative estimates result in empty operating rooms leading to lower occupancy and throughput. Overall, unbiased, and accurate estimates will be the prerequisite to achieve more efficient operating theatre sets, with well-demonstrated repercussions on both clinical outcomes and patient experience (Kayis *et al.*, 2012). This last point is extremely important as, in an increasingly competitive area, the hospital experience felt by each client may be strongly penalised by a delay in the call to the block or an unwanted cancellation of their surgery.
This study used data from surgeries performed between 2019 and 2022 within the orthopaedic specialty of Hospital da Luz of Lisbon, a private healthcare reference entity in Portugal. The main objective of this study was to build a prediction model that can be applied as an autonomous tool to provide more accurate estimates to service coordinators and block managers in charge of planning and sequencing surgeries to ensure their operational efficiency and, secondarily, to identify which preoperative variables are best related to surgical duration. In this way, and considering the existing literature, linear regression was selected

as the resolution approach to be applied. Three possible scenarios were built (one aggregate model, one model per procedure and one model per surgeon) with the purpose of understanding which of the three reproduces predictions that are closer to those observed. In general, all models outperformed the predictions currently estimated by surgeons, all representing better options than the method currently used. However, it was possible to assess lower RMSE values and mostly higher percentages of accurate predictions for the aggregate model.

## 2. Problem Characterisation

### 2.1 OR characterisation

The OR represents an organic and functional unit of a hospital that integrates physical, human, and technical means and is characterised as the point of convergence of most services and specialties. This structure consists of operating rooms, disinfection rooms, support rooms, and anaesthetic induction rooms. Hospital da Luz of Lisbon has sixteen operating theatres, one of which is for robotics, an emergency room, and a Lasik room. Each room has a disinfection room as well as an anaesthetic induction room. This unit is open from Monday to Friday from 8am to 10pm and Saturday from 8am to 2pm. Regarding human resources, for each surgery, Hospital da Luz of Lisbon usually provides 6 staff members: surgeon, anaesthesiologist, anaesthesia nurse, circulating nurse, instrumentalist, and auxiliary nurse.

### 2.2 OR planning and scheduling

OR scheduling is the process of scheduling surgeries by surgeon and by room that has as output a detailed calendar, most often weekly, where the slot allocated to each procedure is described. Hospital managers aim to maximise the performance of OR utilisation through a variety of strategic steps. The literature refers to three surgical scheduling strategies: block-scheduling, open-scheduling and modified block-scheduling.
The strategy followed by Hospital da Luz of Lisbon is a surgical programming in block-scheduling since each specialty is assigned a fixed weekly schedule during which it can perform surgical scheduling. Each specialty is responsible for, together with the surgeons, carrying out its weekly planning considering the slots allocated to it.

### 2.1 Constraints on OR planning and scheduling

In most hospitals OR availability is limited and so there is a strong emphasis on scheduling as many cases as is feasible, safe, and cost-effective. However, while some surgeries have a relatively predictable duration others can have significant variability in their duration (Denton *et al*., 2007). In addition, some surgeons do not work exclusively in one hospital, which sometimes limits the time available to schedule their surgeries.

At Hospital da Luz of Lisbon, and regarding the orthopaedic specialty, surgeons estimate the duration of the surgery they propose to perform based on their experience. There are few tools available to assist in the calculation of surgical time forecasts. However, this is a limitation that can easily be overcome, given that the hospital has rigorous and compiled management support data on the OR activity with which it is possible to work.
Although this is a daily constraint and known by all parties involved, there is still no strict control over the time predicted by surgeons and the time observed, which makes it difficult to implement measures to alert surgeons to the imperative need to minimize these deviations. Figure 1 shows the comparison between the predicted duration of surgery estimated by orthopaedic surgeons at the Hospital da Luz of Lisbon and its actual duration in the period between 1 January 2019 and 31 May 2022, data provided by the hospital following the positive opinion of the respective research and ethics committees and which will be used to develop this study. The red line shows the ideal scenario in which the estimates are always correct. It is possible to observe in the graph that observations tend to be located on the left side of the line. This reflects the obvious overestimation in the surgeons' forecast regarding the actual surgical times, which may encourage periods of OR downtime.

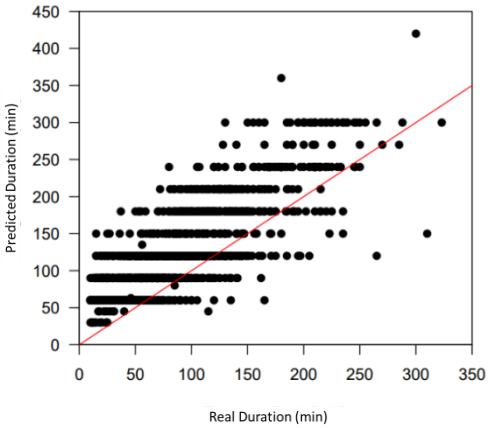


Figure 1 I Comparison between the predicted surgical duration estimated by surgeons and the actual duration, for the period between 1 January 2019 and 31 May 2022.

## 3. Literature review

### 3.1 OR planning and scheduling

Within the scope of surgical time prediction, several approaches have been proposed. These include statistical models such like linear regression (LR) applied by Strum *et al*. (2000), Eijkemans J. *et al*. (2010), Kayis *et al*. (2012) and Edelman *et al*. (2017) or machine learning algorithms used by Tuwatananurak *et al*. (2019), Bartek *et al*. (2019), Zhao *et al*. (2019) and Abbou

*et al*. (2022).
Furthermore, in relation to the independent variables that present a greater relationship with the dependent variable, Ng *et al*. (2017) concludes that the variable with the highest weight in the model was related to procedures. However, the location, patient's class, surgeon, type of anaesthesia and gender of the patient all contribute significantly. Time of day seems to influence the performance of the models contrary to the day of the week. The month seems to introduce noise to the models, leading to a reduction in the performance of the test set. On the other hand, Eijkemans J. *et al*. (2010) finds out that factors related to surgery and team have greater predictive power. Kayis *et al*. (2012) concludes that operational factors (order of surgery, OR allocation surgical team) are promising in improving the predictability of surgery.

### 3.2 Selection of resolution approach

Most of the articles mentioned in the literature analyse statistical models, namely linear regression, as an approach to solve the problem of surgical time prediction in hospital environment. This method is not very flexible since it can only generate linear functions (lines or planes) unlike non-linear approaches that avoid the assumption of a particular functional form for f, allowing a more accurate adjustment to a wide range of possible forms for f.
When inference is the goal, there are clear advantages in using simple and relatively flexible statistical learning methods. In some scenarios, the interest is only in predictability and not in the interpretation of the predictive model. In these cases, it will be better to use the most flexible model possible. Interestingly, this is not always the case. More accurate predictions can be obtained using a less flexible method. This phenomenon, which may seem counterintuitive at first glance, has to do with the potential for overfitting in very flexible methods, an undesirable situation because the fit obtained may not produce accurate estimates of the response when new observations that were not part of the original dataset are used (Hastie *et al*., 2021).
The aim of this dissertation, as already mentioned, is to develop a model that produces accurate estimates that outperform the prediction method for surgery times currently used in Hospital da Luz of Lisbon. To this end, it is necessary to identify which possible preoperative variables can be related to this duration. Thus, the selected resolution approach will have to enable accurate predictions to be achieved, allowing the model to be interpreted.
For this reason, the linear regression method was selected.

## 4. Research Methodology

For each of the models, the methodology adopted follows the following steps: information selection, data preparation, modelling, and results.

### 4.1 Information Selection

Before establishing the prediction model using the historical data, a cleanup action was performed to eliminate data records considered "invalid". These include incomplete records, records relating to CPTs with less than 20 records to ensure the predictive potential of each procedure and records of surgery time with durations of less than 10 minutes because they may represent recording errors, hardly reflecting durations associated to orthopedic procedures. Elective and urgent surgeries were considered as part of the sample, given that the scope of this master's dissertation does not require a distinction between the two types. Observations recorded during the 3rd and 4th quarter of 2020 periods that reflected significant impacts of the pandemic on surgical durations were also disregarded.
The data were subsequently divided into two samples: the training sample containing records from the years 2019, 2020 and 2021 and the test sample integrating observations recorded during the first half of 2022.
This resulted in a training sample of 3312 surgeries recorded in the period between 1 January 2019 and 31 December 2021 (excluding the 3rd and 4th quarters of 2020) and a test sample of 713 surgeries referring to the first half of 2022. The sample used to train the aggregate model includes 41 distinct procedures performed by 33 surgeons while the test sample features 34 distinct procedures performed by 27 surgeons.

### 4.1 Data Preparation

This step is characterised by a set of validations required to ensure the success of the least squares approach, i.e., to ensure that this method reproduces the best linear and unbiased coefficients (Hosseini *et al*., 2015). The assumptions should be validated for all samples that are at the origin of the design of each model. However, in this section we will only exemplify the steps for the aggregate scenario.
In a first stage, and to ensure the best performance of the model to be developed, it should be ensured that the independent variable, in this case the real duration of surgeries, presents a normal distribution. To assure this assumption, the logarithm was applied to this variable.
For the development of the proposed model only the pre-operatively available variables described in table 1 were used. The "mean_duration" was the only variable added to the data shared by the hospital and it translates into a calculated field presenting the mean duration per procedure and per surgeon. This factor was added as it is mentioned in the literature as having predictive potential (Bartek *et al*., 2019). Currently, the average duration is still the prediction used by many hospitals.

Table 1 I Description of the variables available pre-operatively.

| Independent Variables (Factors) | Type | Levels |
|---|---|---|
| month | Nominal | 12 |

| | | |
|---|---|---|
| week_day | Nominal | 5 |
| working_day | Nominal | 2 |
| shift | Nominal | 2 |
| room | Nominal | 14 |
| age_patient | Numeric | - |
| gender_patient | Nominal | 2 |
| procedure_code | Nominal | 41 |
| main_surgeon | Nominal | 33 |
| ambulatory | Nominal | 2 |
| surgery_priority | Nominal | 2 |
| first_surgery_day | Nominal | 2 |
| mean_duration | Numeric | - |
| predicted_duration | Numeric | - |

For each of the scenarios to be tested, the selection of variables to be included in each model considered the validation of the following assumptions:
1) The independent variables are uncorrelated: assessed by Pearson's correlation coefficient.
2) There is no multicollinearity among independent variables: assessed by the VIF (variance inflation factor) metric.
3) The categorical variables are statistically significant: assessed by the p-value resulting from the analysis of variance (ANOVA) which allows for the analysis of the differences between the means referring to the various categories of a group.
The validation of each assumption was performed using the statistical analysis software R, version 4.2.1. The first assumption to be validated applied to the numerical variables ("age_patient", "mean_duration" and "predicted_duration"). With the help of R, it was possible to calculate the correlation coefficient between each pair of variables and confirm the high correlation between the variables "mean_duration" and "predicted_duration", which present a correlation coefficient of 0.89. Thus, it was necessary to disregard one of the variables. The choice of the variable to be rejected is arbitrary, however, the variable removed was "predicted_duration" as it adds the least information to the model and presents a lower accuracy, representing precisely the variable that is intended to be optimised.
To evaluate multicollinearity, it is frequent to use the VIF which evaluates how much the variance of an estimated regression coefficient increases if its variables are correlated. A VIF between 5 and 10 indicates high correlation, which can be problematic. If the VIF presents a value greater than 10, it can be assumed that the regression coefficients are poorly estimated due to multicollinearity. This metric can be applied considering numerical variables and categorical variables that do not present more than two levels. Since all considered variables presented low VIF values, none was disregarded.
For the categorical variables with more than two levels, the ANOVA was performed and resulted in the removal of the variable “month”.
It is from the aggregate data sample that the samples per procedure and surgeon can be created. However, it is necessary to ensure that the training samples are large enough to contribute to a correct prediction. Thus, to obtain data per surgeon, a threshold of 180 observations was set, below which the sample would no longer be able to produce reliable predictions. Therefore, in a first stage, procedures with a number of observations higher than 180 were selected, excluding those with a lower number of records. Each procedure will correspond to a distinct data sample that will allow the development of each model. The same logic was used to obtain data per surgeon, i.e., procedures with a frequency higher than 180 were selected and the remaining procedures were excluded. Thus, 9 distinct models were developed: one aggregate model, four models per procedure and four models per surgeon.

### 4.3 Modelling

After the data preparation step, modelling follows, a phase that begins with the selection of variables that are statistically significant enough to be included in the model. As presented in the theoretical background of the previous chapter, there are three different selection methods: forward selection, backward selection, and mixed selection. As n is considerably greater than p, it follows that the backward selection mechanism can be used.
For the specific aggregate scenario, the variables eliminated were firstly the "week_day" and secondly the "shift", obtaining the model composed of the following variables: "working_day", "age_patient", "gender_patient", "room", "main_surgeon", "procedure_code", "surgery_priority", "first_surgery_day", "ambulatory" and "mean_duration".
In addition to the general assumptions about the correct model specification, it is important for the whole development that there are: (1) non-linear relationships between the independent variables and the dependent variable, (2) that the model errors exhibit normal distribution with zero mean value, (3) that the model errors have constant variance and (4) that there are no outliers that could influence the model results. All these four assumptions can be validated by analysing the graphs in figure 2 (respectively, from left to right and from top to bottom). The "Residuals vs Fitted" chart assesses the linearity of the independent variables with respect to the dependent variable. If the residuals are distributed around a horizontal line with no distinct patterns, this is a good indication that there are no non-linear relationships between variables. If we analyse this graph for the aggregate scenario, we can see that the residuals are almost symmetrically distributed around a horizontal line, which confirms that the model does not exhibit non-linear relationships. The "Normal Q-Q" chart shows whether the residuals are normally distributed or not. If the residuals are normally distributed, they will follow a straight line without large deviations. Analysing this chart for the aggregate scenario, it is confirmed that despite a slight deviation at the tips, the residuals are well aligned on the dashed straight line and are therefore normally distributed. The "Scale-

Location" chart shows whether the residuals are equally distributed along the predictor intervals. This is how the assumption of constant variance (homoscedasticity) can be verified. One way to check this assumption is to analyse if there is a horizontal line with equally (randomly) dispersed points. So, it is possible to conclude from this graphic for the aggregate case that the variance of the residuals remains constant. The purpose of the "Residuals vs Leverage" chart is to detect outliers that may represent influential cases in the linear regression analysis. One should look for values in the upper or lower right corner and watch for cases outside the line representing Cook's distance (cases with high Cook's distance scores) which represent influential cases for the regression results, that is, if we exclude them, the regression results are altered. Observations 39 and 254 are close to the boundary representing Cook's distance, however, they do not exceed it and therefore, despite being outliers, do not represent influential cases that have to be excluded from the model.

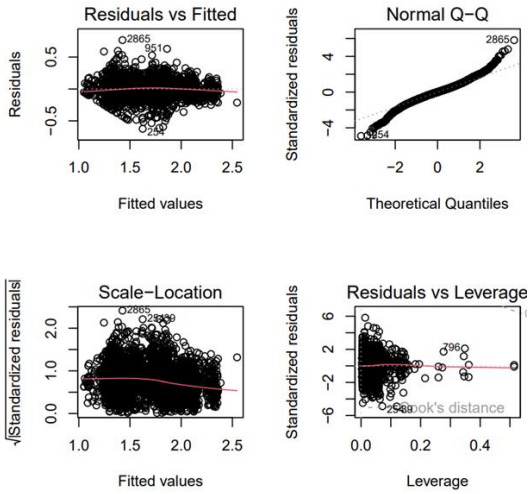


Figure 2 I Residuals Analysis - Aggregate Scenario.

## 5. Results

### 5.1 Results for the three scenarios

In a first stage, the RMSE was determined for each model developed and for the predictions estimated by the surgeons and concluded that all models present more accurate results (lower values of RMSE) than the surgeons' predictions.

In relation to the aggregate model, it is possible to conclude through its adjusted $R^2$ value that it explains 84.29% of the variance of the dependent variable from the independent variables included in the model. This value is much lower for the models by procedure, showing that the independent variables included in the latter do not seem to have a great explanatory power regarding the variance of the response variable. On the other hand, the models per surgeon presented high adjusted $R^2$ values, even exceeding the value of the aggregate model for surgeon 1 and 3. Only the model concerning surgeon 2 originated a lower adjusted R2. Thus, it is possible to conclude that the aggregate model and the models per surgeon tend to better explain the variance of the surgical duration.

To quantify the improvements in the prediction of surgical durations made possible by the models, 3 levels of classification were distinguished:

- Exact: predictions that are included in a certain tolerance range.
- Overestimated: predictions that are higher than the defined tolerance.
- Underestimated: predictions lower than the defined tolerance.

Ideally, a model should provide a higher percentage of exact predictions and overestimate more than it underestimates (Zhao *et al*., 2019). Overestimation will be preferable because, although it may lead to lower block occupancy, it will not contribute to possible cancellations or postponements, both situations that can represent significant weights on patient health and the level of service that is expected from a private healthcare institution. However, this trade-off between block vacancy time resulting from overestimations and the time extra than expected should be strictly monitored so that the opportunity cost of one does not outweigh the other.

Figure 3 demonstrates the percentage of predictions made by surgeons considered as accurate, overestimations and underestimations considering a tolerance interval of 10%, 20% or 30% for surgeries lasting more than 100 minutes and a tolerance interval of 10, 20 or 30 minutes for surgeries lasting less than 100 minutes. As expected, as the tolerance interval increases more predictions are classified as accurate. However, it should be noted that, for a tolerance considered high (30% or 30 minutes), the percentage of overestimated cases is still considerable (63%), leading us to conclude that surgeons, in most cases, overestimate far beyond the actual duration of the surgeries they intend to perform.

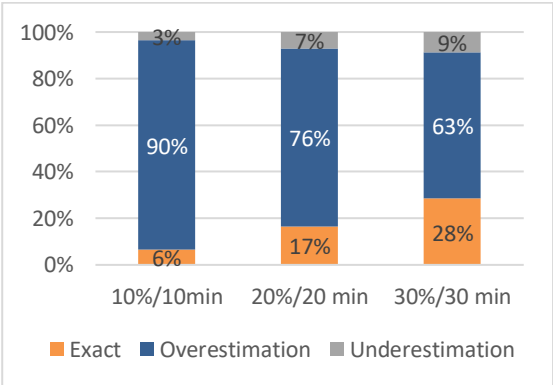


Figure 3 I Graph of the classification of the predictions made by the surgeons - Test Sample.

By analysing the graph in figure 4 regarding the tolerance interval of 10 min for surgeries lasting less than 100 minutes and 10% for surgeries lasting

more than 100 minutes it was possible to conclude, once again, that all models show significant improvements in prediction accuracy when compared to the surgeons' predictions for the same tolerance interval. The aggregate model accurately predicts 58% of the time and overestimates more than it underestimates. The models per procedure accurately predict from a minimum of 28% to a maximum of 74% of cases underestimating more than overestimating. Regarding surgeon models, these accurately predict between 37% and 74% of cases, overestimating more than underestimating.

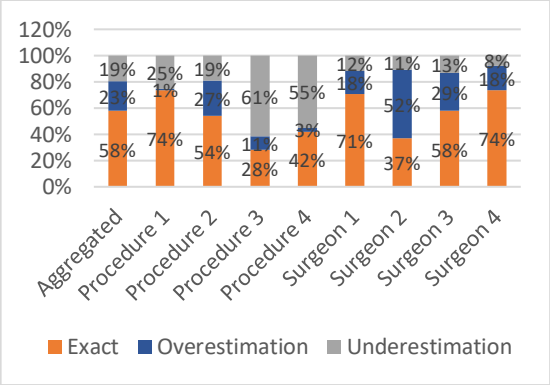


Figure 4 I Classification chart of the predictions made by the models considering a tolerance interval of 10 minutes for surgeries lasting less than 100 minutes and 10% for surgeries lasting more than 100 minutes.

The analysis was repeated for other tolerance ranges. Naturally, as this range increases the greater the percentage of accuracy presented by the models.
If the hospital chooses to use a model as a tool to support the calculation of surgical time forecasts, it should define the tolerance allowed and then choose the model to be used. This tolerance should ideally be communicated to and considered by the person who subsequently carries out the weekly occupancy planning of the OR.
However, the analysis carried out does not allow us to deduce, which model effectively reproduces the most accurate forecasts. To do so, it is necessary to compare the models individually with each other, comparing samples similar in size and information, an analysis that is reproduced in the following sections. However, to conclude whether the model per procedure outperforms the estimates of the model per surgeon or the opposite, it would be necessary to filter the results of both models simultaneously, that is, the models per procedure would be filtered by surgeon and compared with the results of the models per surgeon filtered for each procedure. This would result in very small data samples that would provide conclusions with little robustness. For this reason, it was decided not to perform this comparison.

### 5.2 Comparison between the aggregate scenario and the scenario per procedure

To be able to state which model had the best performance, in a first stage, the performance of the aggregated model was compared with that of the model per procedure. To this end, the results of the aggregated model were filtered by procedure and compared with those of the model relative to the corresponding procedure. This analysis was performed for the four selected procedures using, in a first stage, the RMSE metric (table 2).

Table 2 I Comparison of the RMSE of the aggregate model calculated for each procedure and the RMSE of each model by procedure.

| Procedure | Aggregate RMSE | Procedure RMSE |
|---|---|---|
| Procedure 1 | 11.8 | 14.0 |
| Procedure 2 | 15.5 | 15.4 |
| Procedure 3 | 21.8 | 27.5 |
| Procedure 4 | 20.5 | 18.7 |

Table 2 shows little significant differences between the RMSE of the aggregate model filtered by procedure and the RMSE of the model by procedure. However, for procedure 1 and 3, the aggregate model has a lower RMSE, a value that translates into a higher percentage of accuracy than the per-procedure model, with this difference being particularly notable in procedure 3. For procedure 2, the RMSE of the per-procedure model is slightly lower than that of the aggregate model, a value also mirrored in the accuracy analysis where the percentage is higher for the per-procedure model than for the aggregate model. Regarding procedure 4, the RMSE of the aggregate model is higher than that of the model by procedure. However, from figure 5 it can be seen that the percentage of accuracy remains higher for the aggregate model when compared to the per procedure model. A possible explanation for this fact may lie in the low adjusted $R^2$ values for the per-procedure models relative to the aggregate model. These values may be justified by a limitation of the data provided. The codes of the shared procedures are not the standardised and universally used CPTs. The codes of the sample are internal to the hospital and include all procedures and sub-procedures that the surgeon intends to perform. This means that for the same procedure, if the sub-procedure is distinct or extra to the basic surgery, the surgery will present another code, even if in terms of complexity and surgical time it has no influence. This constraint directly impacts the sample size of each procedure available which could be considerably larger and thus contributing to the training of more accurate models per procedure. Thus, it would be more favourable to opt for the aggregate model which seems to better explain the variance in the duration of surgeries, presenting low RMSE values and accuracy percentages sufficiently high to constitute a good alternative to the current estimation model.

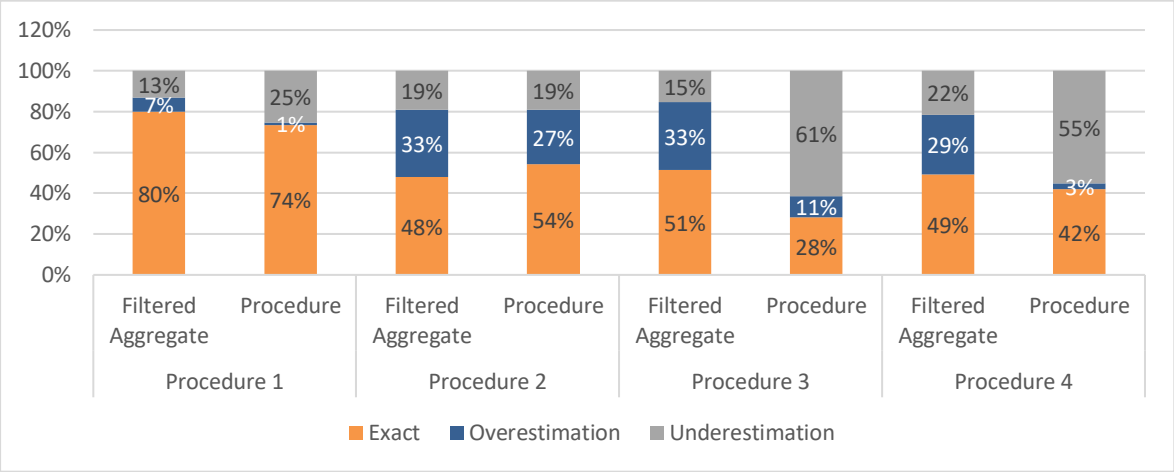


Figure 5 I Graph showing the ranking of the predictions made by the aggregate model filtered by procedure and by the model by procedure considering a tolerance interval of 10 minutes for surgeries lasting less than 100 minutes and 10% for surgeries lasting more than 100 minutes.

### 5.3 Comparison between the aggregate scenario and the scenario per surgeon

As in the previous section, the comparison analysis between the aggregate model and the model per surgeon was performed. The results of the aggregate model were now filtered by surgeon and compared with those of the model relative to the corresponding surgeon. This analysis was performed for the four selected surgeons using also the RMSE metric (table 3).

*Table 3 I Comparison of the RMSE of the aggregate model calculated for each surgeon and the RMSE of each model per surgeon.*

| Surgeon | Aggregate RMSE | Procedure RMSE |
|---|---|---|
| Surgeon 1 | 12.2 | 16.6 |
| Surgeon 2 | 30.5 | 37.1 |
| Surgeon 3 | 11.0 | 13.8 |
| Surgeon 4 | 11.1 | 11.7 |

When comparing the RMSE values presented in table 3 it is possible to see that, once again, there are no significant differences between the two models. Despite this fact, consistently lower values are found for the aggregate model. Even so, the models per surgeon return quite satisfactory adjusted $R^2$ values, with two of them even obtaining higher values than the aggregate model (see table 3). Only surgeon 2 is associated with a lower adjusted $R^2$ value, a behaviour that is subsequently reflected in its higher RMSE and in its accuracy percentage which is significantly lower when compared to the other models (see figure 6). This fact may be explained by the fact that the procedures performed by this surgeon in the test sample present significantly longer mean durations than the procedures recorded for the other surgeons analysed. Strum *et al*. (2000) refer that the absolute variability is expected to be higher for surgeries with high durations, which justifies the higher RMSE and lower percentage of accuracy for this surgeon.

Regarding the percentages of accuracy, these also do not vary much between models, being mostly higher for the aggregate model. However, the models per surgeon show consistently higher percentages of overestimation than of underestimation, while the aggregate model shows the same frequency of overestimation and underestimation, although overestimating to a greater extent.

Thus, it is predicted that the aggregate model will be the best option as an alternative to the method currently used. Even so, the per surgeon models show potential to possibly surpass the aggregate model when trained with a larger sample than the current one. A sample size of 3312 observations was used to train the aggregate model while the per-surgeon models were trained using sample sizes ranging from 227 to 625 observations, a difference which is confirmed to be relevant. This conclusion may be supported by the study of Bartek *et al*., (2019).

### 5.4 Analysis of the relationship between predictors and the dependent variable

The models included operational variables (month, day of the week, working day, shift, room, whether or not the surgery was the first of the day, priority of the surgery), variables representing patient characteristics (age and gender), variables related to procedure characteristics (procedure code, whether it was an outpatient surgery or not, the average duration per procedure and per surgeon and the duration expected by the surgeon) and a variable representing the team (head surgeon).

As expected, the variable "procedure_code" was included in all models and was, therefore, the factor with the highest predictive power, a conclusion that is

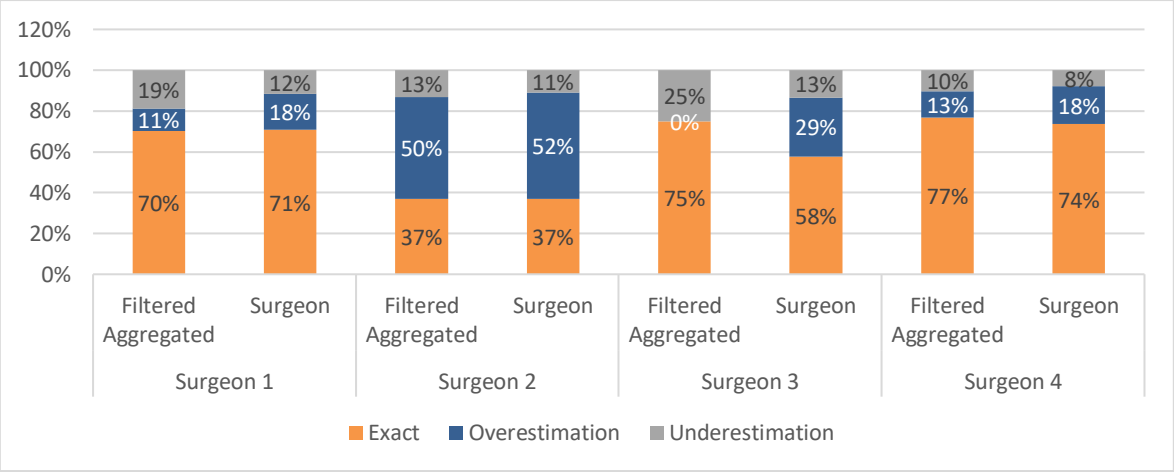


Figure 6 I Graph showing the ranking of the predictions made by the aggregate model filtered by surgeon and by the model by surgeon considering a tolerance interval of 10 minutes for surgeries lasting less than 100 minutes and 10% for surgeries lasting more than 100 minutes.

already mentioned in the literature (Ng *et al.*, 2017). Also, regarding the variables that characterise the procedures, we found that at least one of the durations (mean or expected) was relevant in the construction of the aggregate and per-procedure models, which were not very useful for the models per surgeon. On the other hand, we concluded that patient characteristics also had a significant weight and at least one of them was included in all models, except for surgeon 3. In addition, the main surgeon seemed to have a similar impact on the duration of surgeries, and this variable was selected for all aggregate models and by procedure, except for procedure 2. Finally, we found that among the operational variables, those that showed a greater relationship with the surgical time were the priority of surgery (elective or urgent) and whether the surgery was the first of the day or not. The variables "month" and "week_day" do not seem to influence the performance of the models as is the case in the work conducted by Ng *et al*. (2017). However, in the referenced study the time of day seems to be significant, however, in the present work the opposite is demonstrated with the variable "shift" being considered only in the model regarding procedure 2.

## 6. Discussion

The OR is the most critical and expensive resource of a hospital, representing, on the other hand, also its greatest source of income. Therefore, it is essential that it is managed efficiently since each minute wasted can cause significant loss of income. For an efficient use of the OR, accurate OR occupancy time forecasts are necessary to contribute to a better sequencing of surgeries.

The OR occupancy time is characterized by several phases and presents many factors that may interfere with its efficiency and that make it demanding to plan surgeries in a way that maximizes the efficiency of resources and avoids overestimations or underestimations. These cases may cause, advances (resulting in losses in block occupation), delays or even cancellations that may have negative impacts on patient health as well as on the level of service that is intended to be offered to the client, a critical factor in a private health service unit.

In this context, the aim of this work is to develop predictive models for surgery duration that overcome the estimates currently made by surgeons of the orthopaedic specialty at Hospital da Luz of Lisbon. Since few tools are currently available to support prediction, surgeons predict the surgical time based only on their experience. According to the analysis developed, this results in very low accuracy percentages of only 6% considering a tolerance interval of 10 minutes for surgeries lasting less than 100 minutes and 10% of the real duration for surgeries lasting more than 100 minutes. Although there is no consensus in the literature regarding the definition of surgical duration, this study considers the duration between incision and patient closure.

Within the scope of the optimisation of surgical time prediction, several approaches have been proposed in the literature. However, the selected approach was linear regression because it ensures the best trade-off between flexibility and interpretability.

For this work, we chose to develop three different scenarios: an aggregate model, models per procedure and models per surgeon. The shared data sample includes surgeries performed between the years 2019 and 2022, corresponding to a total of 8937 surgeries only concerning the orthopaedic specialty. After the data selection and preparation phase, this sample was reduced to 3312 surgeries used to train the models and 713 surgeries that made up the test sample. After ensuring the necessary assumptions, nine different models were modelled (one aggregate model, four models per procedure and four models per surgeon).

In general, the developed models produced better predictions than the surgeons. The RMSE of each proved to be considerably lower than the method currently practised, and the percentage of accuracy rose to values between 28% and 74%. These percentages reflect a conservative tolerance range and analyses have been carried out for higher ranges

for which the percentage of accuracy reaches 98% for one of the models. However, to determine which model shows the best performance it was necessary to compare them individually. When comparing the aggregate model with the models by procedure it was possible to conclude through a much higher adjusted $R^2$ value that the aggregate model has a greater explanatory power regarding the variance of the actual duration of surgeries. In addition to this fact, the reduced RMSE values and the high percentage of accuracy make it a valid alternative to the current standard. In a second stage, the aggregate model was compared with the model per surgeon. It was possible to observe consistently lower RMSE values for the aggregated model and slightly variable accuracy percentages between both, but mostly higher for the aggregated model. However, the models per surgeon return quite satisfactory adjusted $R^2$ values, two of them even obtaining higher values than the aggregate model. Thus, everything indicates that the aggregate model represents the best alternative to the predictions currently estimated by surgeons; however, the models per surgeon show potential to possibly outperform the aggregate model when trained with a larger sample than that currently available. It is also considered that the use of this methodology, in addition to contributing to an improvement in OR efficiency, may progressively raise awareness among surgeons of the importance of making estimates as accurate as possible, thus cultivating a culture of co-responsibility in the management of OR time.

Despite the results obtained, there are important limitations that should be considered. The data were collected retrospectively, so no claims can be made about the accuracy of the times recorded by the OR team. On the other hand, and as previously mentioned, the codes per procedure shared do not correspond to the universally used CPTs. The codes used for the analysis are internal to the hospital and indicate all procedures and ancillary procedures to be performed during surgery. It so happens that some of these secondary procedures, having no real impact on the total surgery time, form new combinations of surgical codes and thus, new types of surgery. This fact, had a direct impact on the number of observations per type of procedure that could be higher in some cases and, consequently, originate better trained models. Effectively, it was possible to deduce from the results that the models per surgeon have the potential to possibly outperform the aggregate model; however, the models were trained with a substantially smaller sample than the aggregate model, and this may be one of their limitations. Additionally, it was not possible to quickly obtain data regarding some of the variables that the literature studied refers to as having predictive power, such as patient comorbidities or ASA risk class. Finally, this study was carried out only for the orthopaedic specialty of Hospital da Luz of Lisbon, and it is probably not generalizable to other specialties or facilities.

Finally, although the effective duration of surgery is possibly the most critical factor to be considered, the optimisation of OR occupancy depends in a similar way on the monitoring of the remaining phases that integrate the operative process such as anaesthesia time and especially turnover time. The latter represents the time between the exit of a patient from the operating room and the entrance of a new patient and when the materials are removed from the operating room, cleaned, and prepared for the beginning of the next surgery. In most cases, according to the literature, there is often a lack of operational flow or a lack of standardised procedures that minimise the time associated with this stage.

## 7. References


Abbou, B., Tal, O., Frenkel, G., Rubin, R., & Rappoport, N. (2022). *Optimizing Operation Room Utilization — A Prediction Model*. 1–13.

Bartek, M. A., Saxena, R. C., Solomon, S., Fong, C. T., Behara, L. D., Venigandla, R., Velagapudi, K., Lang, J. D., & Nair, B. G. (2019). Improving Operating Room Efficiency: Machine Learning Approach to Predict Case-Time Duration. *Journal of the American College of Surgeons*, *229*(4), 346-354.e3. https://doi.org/10.1016/j.jamcollsurg.2019.05.029

Denton, B., Viapiano, J., & Vogl, A. (2007). Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science*, *10*(1), 13–24. https://doi.org/10.1007/s10729-006-9005-4

Edelman, E. R., van Kuijk, S. M. J., Hamaekers, A. E. W., de Korte, M. J. M., van Merode, G. G., & Buhre, W. F. F. A. (2017). Improving the prediction of total surgical procedure time using linear regression modeling. *Frontiers in Medicine*, *4*(JUN), 1–5. https://doi.org/10.3389/fmed.2017.00085

Hastie, T., Tibshirani, R., James, G., & Witten, D. (2021). An introduction to statistical learning (2nd ed.). *Springer Texts*, *102*, 618.

Hosseini, N., Sir, M. Y., Jankowski, C. J., & Pasupathy, K. S. (2015). Surgical Duration Estimation via Data Mining and Predictive Modeling: A Case Study. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2015*, 640–648.

Eijkemans J. *et al*. (2010). Predicting the unpredictable. *Heart Rhythm et Al.*, *7*(1), 72–73. https://doi.org/10.1016/j.hrthm.2009.10.001

Kayis, E., Wang, H., Patel, M., Ms, T. G., Jain, S., Ramamurthi, R. J., Santos, C., Singhal, S., Suermondt, J., & Sylvester, K. (2012). Improving Prediction of Surgery Duration using Operational and Temporal Factors HP Labs , Palo Alto , CA ; Lucile Packard Children ' s Hospital , Palo Alto , CA and. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, 456–462.

Lee, D. J., Ding, J., & Guzzo, T. J. (2019). Improving Operating Room Efficiency. *Current Urology Reports*, *20*(6). https://doi.org/10.1007/s11934-019-0895-3

Ng, N., Gabriel, R. A., McAuley, J., Elkan, C., & Lipton, Z. C. (2017). *Predicting Surgery Duration with Neural Heteroscedastic Regression*. *68*, 1–12. http://arxiv.org/abs/1702.05386

Strum, D. P., Sampson, A. R., May, J. H., & Vargas, L. G. (2000). Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology*, *92*(5), 1454–1466. https://doi.org/10.1097/00000542-200005000-00036

Tuwatananurak, J. P., Zadeh, S., Xu, X., Vacanti, J. A., Fulton, W. R., Ehrenfeld, J. M., & Urman, R. D. (2019). Machine Learning Can Improve Estimation of Surgical Case Duration: A Pilot Study. *Journal of Medical Systems*, *43*(3). https://doi.org/10.1007/s10916-019-1160-5

Zhao, B., Waterman, R. S., Urman, R. D., & Gabriel, R. A. (2019). A Machine Learning Approach to Predicting Case Duration for Robot-Assisted Surgery. *Journal of Medical Systems*, *43*(2). https://doi.org/10.1007/s10916-018-1151-y