Simulation and Assessment of a Forecasting Model for Operating Room Management

The case of Hospital da Luz Lisboa

Carolina das Neves Narciso
Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal, carolina.narciso@tecnico.ulisboa.pt

The uncertainty inherent to surgery duration forecast is one of the most challenging issues of managing the operating room (OR). Since the OR holds a significant amount of a hospital’s total costs and revenues, formulating accurate forecasts is a priority to increase the OR efficiency. Anonymised data between January 2016 and August 2019 revealed surgeries performed at Hospital da Luz Lisboa (HLL) with significant deviations between the surgery duration predicted by the surgeon and the real duration. Luz Saúde developed a machine learning (ML) model to forecast the surgery duration with the objective to establish more accurate forecasts. This research arose of a stage of analysis of a possible model’s deployment. The purpose was to understand the risk (i.e., the gain and cost) of deploying the model in daily clinical practice. It was developed an assessment of the quality of the forecasts made by the model in a new dataset never seen before by the model, and it was performed a comparative analysis between the error of the model and those derived from the current method used to predict the duration of surgery (the surgeon’s forecasts). The research also includes the development of a genetic algorithm approach to automate the design of elective surgery schedules; the goal was to assess the impact of using the model in planning the OR schedules. In the end, the results show that adopting Luz Saúde’s model along with the genetic algorithms is an innovative and practical approach for the OR planning.

Keywords: Operating Room Optimization, Surgery Duration Forecast, Surgery Scheduling, Machine Learning

1 INTRODUCTION

The Operating Room (OR) is one of the most critical and demanding healthcare resources within a hospital. In addition to being one of the main cost drivers, the OR is a resource that significantly contributes to the hospital’s revenue [1]. Therefore, the OR management must be focused on system efficiency and optimisation.

The uncertainty associated with surgery duration forecast is one of the most challenging issues in OR management. This topic has been widely discussed in the literature in the last years [2], [3]. Problems associated with under or overutilisation of the OR are undesirable. Therefore, they have become a high priority for the hospital’s administration, which seeks to offer health care with high quality and productivity to its users.

Currently, the primary surgeon’s forecasts, based on his experience, are used in many hospitals to estimate the surgery duration and design the OR schedules. However, the surgeon’s forecasts are often imprecise and result in poor use of the OR.

Given the above problem, hospital administrators explore new strategies directed to efficient and effective planning and scheduling of surgeries. Since 2015, the scientific community has been discovering the potential of ML algorithms in OR management [4]. Different research teams exploit forecasting algorithms based on time series to predict the surgery duration. The forecast made by ML algorithms shows positive results compared to traditional approaches [5]–[7]. Thus, forecasting algorithms have become a promising tool to provide reliable plans of elective surgeries without under or overrunning.

In addition to being important to assess the quality of results from these types of approaches, it is crucial to ascertain the practical significance of the incremental benefit of the forecasts when the algorithms are deployed in clinical practice.

This research was developed in line with the project already initiated by Luz Saúde, one of Portugal’s largest private healthcare groups, in a pilot exercise for HLL. Luz Saúde created a ML model for predicting the duration of surgery based on clinical-surgical administrative anonymised data from HLL. Luz Saúde’s principal goal would be to formulate more accurate forecasts of the surgery duration to optimise OR efficiency. Data between January 2016 and August 2019 revealed that a considerable percentage of surgeries performed at Hospital da Luz Lisboa had significant deviations between the surgery duration predicted by the surgeon and the real duration of the surgery in the OR.

Luz Saúde intends to put its forecasting model into production as a tool for managing the OR. However, it is expected that applying the forecasting model will impact the OR planning. As the OR is an impactful area in the hospital, implementing ML models raises some problems and uncertainties.

This research arose from a stage of analysis of a possible model’s deployment. The purpose was to understand the risk of deploying the model in clinical practice: the gain and cost that the HLL could acquire by using the model. Mainly there were two big goals: assess the model’s deployment in daily clinical practice and analyse the impact that the application of the model’s forecasts would have to optimise the planning of elective surgeries throughout the theatres of the OR.

To assess the model’s deployment in daily practice, the quality of the forecasts was validated in a new dataset never seen before by the model. A prototype scheduling model, based on a genetic algorithm, is proposed to analyse the performance of the model in optimising elective surgery planning in the OR.

The following section contains the case study description. The third section reviews the relevant literature according to the problem of forecasting the surgery duration, highlighting the current state-of-the-art. In the fourth section, the methodology to assess the model and analyse its performance when it is deployed
is described. The evaluation of the quality of the forecasts is developed in the fifth section and the sixth details the genetic algorithm approach proposed to design surgery schedules. Finally, in the last section, the thesis is concluded.

2 HOSPITAL DA LUZ LISBOA’S OPERATING ROOM

Most of the surgeries performed at HLL are scheduled in advance (i.e., by appointment). Ophthalmology, gynaecology-obstetrics, orthopaedics, general surgery, urology and neurosurgery are the specialities with the most significant surgical volume reported by Luz Saúde. They correspond to 80% of activity in the OR. At HLL, OR is distributed primarily by the several specialities based on the volume of their activities. Also, Luz Saúde mentions that the OR planning is based on a “block-bookings” type of functioning. The operating theatres and their available time are divided into slots assigned to each surgery team and surgeon.

Luz Saúde reports that the OR schedules are developed by the surgery block’s team every week. These weekly meetings are time-consuming, focusing on details that could be avoided (e.g., if surgery duration forecasts were more accurate, replanning times would not be necessary). Despite all efforts to date, meaningful deviations between the surgery duration predicted by the principal surgeon and the observed duration in the OR were recorded for a considerable percentage of surgeries performed at HLL between January 2016 and August 2019 (Figure 1). As shown in Figure 2, on average, 45% of the surgeries in the six principal surgical specialities present a deviation of duration greater than 25% compared with the surgeon’s forecast.

![Figure 1 - Average duration of surgery estimated by the surgeon and average duration of surgery in the operating room for the specialties with the highest surgery rate at HLL. Dataset: January 2016 – August 2019.](image1)

![Figure 2 - Percentage of surgeries with deviation in duration greater than 25% compared to the surgeon’s estimate. Dataset: January 2016 – August 2019.](image2)

One of the main challenges identified whilst trying to solve the previous challenge relates to the surgical procedure’s duration. Often, when recording the data, an estimate of the surgery duration is performed usually by the surgeon. However, different physicians report different times when they are estimating. According to Luz Saúde, most surgeons only consider the operative timespan (the time from skin incision to skin closure) and a small proportion of surgeons make an estimate regarding the total duration of the OR occupation (from the patient’s entry into the OR to his transfer to the Post-Anaesthetic Care Unit).

The more accurate the surgery duration forecast, the more efficient is the resources management, the patients’ overall satisfaction, as well as the overall OR profitability. Therefore, the Strategy and Business Analytics of Luz Saúde, in partnership with the CAA of Fidelidade, developed a ML forecasting model that predicts the duration of surgery in the OR.

3 LITERATURE REVIEW

In the last decade, a wide range of ML technologies, especially forecasting algorithms based on time series, has been applied in different studies to predict the surgery duration and assist in OR planning [4]. Linear Regression [8], [9], Neural Networks [10], [11], Random Forests [6], [12] and Gradient Boosted Regression Trees [5], [7] are some examples of the most explored methods in the field. ML approaches demonstrate considerable improvements in forecasting accuracy compared to traditional methods. For example, the forecast accuracy of the Random Forest algorithm increases by 30% to 35% compared to traditional methods [12]. However, the forecast error in forecasting models is still significant despite research efforts to date. For example, Jiao et al. (2020) achieved the model’s forecasts with mean absolute errors (MAEs) of 22.1 minutes for simple Decision Tree, 19.5 minutes for Gradient Boosted Decision Tree and 19.6 minutes for Random Forest for the test data.

4 THESIS STATEMENT

Firstly, because Luz Saúde and CAA did not provide the code with data transformations, it was required to reproduce them to get the features under the same training conditions and then be able to apply the model. To reverify the performance of the forecasting model for surgery duration, the quality of the forecasts was validated on a new dataset never seen before by the model. The new dataset consisted of more than 13,000 elective surgeries performed at the HLL between September 2019 and December 2020 and was designated as the validation dataset. Additionally, to analyse the model’s performance in the new dataset, it was necessary to reproduce the test under the old data used by Luz Saúde. Luz Saúde provided the training/testing dataset, but there was no information about which observations were used by Luz Saúde and the CAA in training and testing. Therefore, the test was reproduced in this research using the training/testing dataset as a whole, considering that the obtained outcome would be an optimistic forecast once that much of the data had been used for model training. The training/testing dataset consisted of more than 43,000 elective surgeries performed at the HLL between...
January 2016 and August 2019. Luz Saúde provided the two datasets.
As a result, the quality of the model’s forecasts was measured at several levels, such as non-outliers versus outliers, shorter versus longer operating time, relation (excess or defect) between forecasts and observed surgery duration in the OR, main surgical specialties, and before versus during the Covid pandemic.
To understand the magnitude of the errors of the model, for each of the mentioned levels, a critical comparison between the errors of the forecasts produced by the model and those made by the primary surgeon was performed, faced with real data (i.e., the observed surgery duration in the OR of HLL). The forecasts of the surgeon are the current method used by HLL to manage surgeries in the OR.
In addition to the error analysis, possible disadvantages in the data preparation process and training of Luz Saúde’s model are described, and other ways of approaching the data, which we anticipate will help minimise forecast error, were suggested. Scheduling surgeries is a complex and challenging task in OR management. It is a time-consuming process associated with complex combinatorial optimisation of constraints, so it is challenging to find optimal solutions for the scheduling problem. Furthermore, the surgery duration forecast error has an additional impact on the surgery schedule. A genetic algorithm approach to design surgery schedules was trained in Python, considering a limited number of physical constraints associated with the OR. The working day schedules from the September 2019 – December 2020 dataset were designed by the genetic algorithm program. For each day, three schedules were designed with consideration of the following:
1. surgery duration forecasts of the Luz Saúde’s model
2. surgery duration forecasts of the primary surgeon
3. real data (i.e., the observed surgery duration)
After this, a comparison is made between the schedule’s fitness values. As a result, an analysis of the performance of the model when it is deployed for developing OR schedule was performed.

4.1 Luz Saúde’s Forecasting Model for Surgery Duration
Luz Saúde’s forecasting model for surgery duration was provided as a pickle file. The model consists of a Random Forest trained over a dataset with a total of 26 input features. The features include characteristics of the patient, the surgical procedure(s) and the surgical team. From the training/testing dataset (January 2016 to August 2019), 80% of the data was used for the model’s training and the remaining 20% was used for testing.
Table 1 presents the assessment metrics obtained by Luz Saúde and the CAA for the training and testing datasets and the cross-validation training strategy. Based on the metric measurements obtained, Luz Saúde considers the results of the model to be positive and following state of the art.

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Test</th>
<th>Cross - Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average error</td>
<td>0.2297</td>
<td>0.2606</td>
<td>0.2358</td>
</tr>
<tr>
<td>Accuracy</td>
<td>94.60%</td>
<td>93.89%</td>
<td>94.46%</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.785</td>
<td>0.723</td>
<td>0.722 (+/- 0.014)</td>
</tr>
<tr>
<td>Mean Square Error (MSE)</td>
<td>0.089</td>
<td>0.114</td>
<td>0.094</td>
</tr>
<tr>
<td>Mean Average Percentage Error (MAPE)</td>
<td>5.40</td>
<td>6.11</td>
<td>5.539</td>
</tr>
</tbody>
</table>

5 EVALUATION OF THE MODEL AND OUTCOME ANALYSIS
For the training/testing of the model, Luz Saúde and CAA used 42,974 surgeries from the January 2016 – August 2019 dataset, after excluding 284 surgeries identified as outliers through the Interquartile Range (IQR) method. Applying the IQR method on the validation dataset (September 2019 – December 2020), 13,573 surgeries are identified as non-outliers and 15 as outliers. For surgeries considered non-outliers, the model predicts surgery duration with a Mean Absolute Error (MAE) of 25.66 minutes for the January 2016 – August 2019 dataset; considering the September 2019 – December 2020 dataset, the MAE of the model’s forecasts is 31.07 minutes (Figure 3). Thus, despite not being perfect, the MAE for the forecasts of Luz Saúde’s model agrees with the study of Jiao et al. (2020).

MAE for two forecasts – model and surgeon – concerning the observed duration of surgery in the OR is practically the same for the scenarios with or without exclusion of outliers (Figure 3). For the January 2016 – August 2019 dataset, the model’s MAE differs less than 1 minute from the surgeon’s MAE; for the September 2019 – December 2020 dataset, the difference is approximately 5 minutes. The advantage of adopting the model to predict the surgery duration over the method currently used to predict the surgery duration and plan the OR lies in practical terms, since the model’s forecast is free from human labour, and therefore from all influencing factors and errors associated with human work.

5.1 Non-outliers versus Outliers
The outlier surgeries consist of long and short surgeries and surgeries with recording errors (an analysis of these latest
surgeries can be found in the thesis). They correspond to 0.66% of the training/testing dataset surgeries. For Data Science, excluding 1% of observations to develop a ML model is perfectly reasonable. However, it is necessary to explore the impact in practical terms that discarding these observations on the model’s training has on its adoption in the field since these surgeries are part of reality and dataset.

Considering those outliers, the forecast of the surgery duration by the model is quite different from the observed surgery duration in the OR. Furthermore, as we can see in Figure 3, the MAE of the model increases drastically for observations considered outliers compared to non-outliers. The model’s MAE is 511.42 minutes for the training/testing dataset (Figure 3). As a result, the model’s adoption to predict the duration of these surgeries is neither accurate nor safe.

Implementing simple univariate statistics like standard deviation (SD) to identify and remove outliers from a data sample was evaluated. The common cut-off point of three SD from the mean was used to identify outliers (i.e., 99.7% of the sample).

Implementing the SD method, 43,109 surgeries are considered non-outliers and 149 surgeries are identified as outliers for the January 2016 – August 2019 dataset; 13,566 surgeries are identified as non-outliers and eight as outliers for the September 2019 – December 2020 dataset.

As shown in Figure 4, the MAE metric for the two datasets does not show any notable differences between the outliers identified by the IQR method or the SD method. With these results, choosing either method would be legitimate. Therefore, the choice of the Luz Saúde of the IQR method is safe in the great majority of situations.

![Figure 4 - MAE in minutes of the surgeon and model forecast concerning the observed duration of surgery in the OR for the two datasets January 2016 – August 2019 and September 2019 – December 2020. A: Use of the IQR method. B: Use of the SD method. Surgeon forecast metrics against real data are represented in the darkest colour. Luz Saúde’s model forecast metrics against real data are marked in a lighter colour.](image)

5.2 Shorter versus Longer Operating Time

For non-outlier surgeries, the two datasets under analysis (January 2016 – August 2019 and September 2019 – December 2020) were divided into two subsets. The median duration of surgery in the OR was the statistical metric used for the division: 75 minutes for the January 2016 – August 2019 dataset and 82 minutes for the September 2019 – December 2020 dataset.

For the January 2016 – August 2019 dataset, 21,583 surgeries occupied the OR between 13 and 75 minutes, and 21,391 used the OR between 76 and 459 minutes (i.e., 7 hours and 39 minutes). The model’s forecast concerning the observed surgery duration in the OR presents a MAE of 14.06 minutes for surgeries with duration below the median and 37.37 minutes for surgeries with duration above the median. For the December 2019 – December 2020 dataset, 6,792 surgeries occupied the OR between 11 and 82 minutes, and 6,767 used this resource between 83 and 585 minutes (i.e., 9 hours and 45 minutes). The model’s forecast presents a MAE of 14.58 minutes for surgeries with duration below the median and 47.62 minutes for surgeries above the median (Table 2).

Table 2 - Range of surgery duration in the OR, number of surgeries and MAE of the model in reference to the duration of surgery in the OR for the two subsets of non-outlier surgeries: surgeries with duration below the median and surgeries with duration above the median. January 2016 – August 2019 and September 2019 – December 2020 datasets. Use of the IQR method.

<table>
<thead>
<tr>
<th>Duration Interval</th>
<th>January 2016 – August 2019</th>
<th>September 2019 – December 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below median</td>
<td>Above median</td>
<td>Below median</td>
</tr>
<tr>
<td>% of non-outliers</td>
<td>50.2%</td>
<td>49.8%</td>
</tr>
<tr>
<td>MAE model vs real</td>
<td>14.06 min</td>
<td>37.37 min</td>
</tr>
</tbody>
</table>

As mentioned, short-term and long-term surgeries are among the outliers. For the outliers of the January 2016 – August 2019 dataset, the model predicts surgery durations with a MAE of 31.23 minutes for short surgeries and 384.67 minutes (i.e., 6 hours and 25 minutes) for long surgeries. Regarding the outliers of the September 2019 – December 2020 dataset, the model’s forecast has a MAE of 23.65 minutes for short surgeries and 503.65 minutes (i.e., 8 hours and 24 minutes) for long surgeries (Figure 5).

The MAE of the model’s forecast concerning the real surgery duration increases with time. In general, for surgeries lasting between 10 and 80 minutes, the model’s forecasts present a MAE of 14 minutes; for surgeries lasting between 80 minutes and 480/540 minutes (i.e., 8/9 hours), the forecasts have MAE in the order of 40 minutes (Figure 5). For short-term surgeries, the MAE of the model’s forecast is in the order of 25 minutes (Figure 5), an error of twice the surgery duration in the OR. For surgeries lasting more than 9 hours, the model’s forecast has a MAE of 8 hours (Figure 5), an error whose value in hours is quite disturbing for a hospital department.

Simulation and Assessment of a Forecasting Model for Operating Room Management

- 4 -
Based on these results, short surgeries lasting less than 10 minutes and very long surgeries lasting more than 9 hours will be the ones that cause the most consequences and make the planning of the OR more difficult with the application of the model, since these surgeries are also the ones with the highest MAPE, as we can see in Figure 5. Indeed, the implementation of the model to predict the duration of short and long surgery and, subsequently, to prepare the OR schedules will have a considerable impact that will undoubtedly result in notable losses. It is important to point those short surgeries and very long surgeries represent together less than 1% of the dataset. Hence, punctual manual corrections are possible to be carried out on the OR schedules.

Furthermore, it should also be noted that MAE of the surgeon’s forecasts does not show notable differences compared to the MAE of the model’s forecasts for each group of surgeries mentioned (Figure 5). The most significant difference is for long-term surgeries.

5.3 Relation (excess or defect) between model and surgeon estimations with observed surgery duration in the OR

As shown in Figure 6, for non-outlier surgeries, there is no direct relationship in terms of overestimation or underestimation for the two datasets: approximately 50% of the surgeries have a model’s forecast above the observed surgery duration, and 50% have a forecast downwards. This result might introduce some bias in the final results. If the model’s forecasts were essentially by excess, the OR planning using the model would not be ideal, but it would still be possible to execute it: there would be time left over. Unfortunately, forecasts by default provide little confidence in applying the model in the OR planning: surgeries will require more time than predicted. Therefore, either the OR would have to close after closing time or surgeries would have to be cancelled and scheduled for the following days.

Furthermore, it is visible in Figure 7 that the surgeon’s forecasts are more discrepant in relation to the real duration than the model’s forecasts for non-outlier surgeries, especially for surgeries with forecasts by excess. For the January 2016 – August 2019 dataset, almost all surgeries show a deviation between the surgeon’s forecast and the real duration that exceeds 100% of the real surgery duration, while for the model forecast, only half of the surgeries have deviation greater than 100% (Figure 7, E). For the September 2019 – December 2020 dataset, the relationship is analogous: for the surgeon’s estimation, all surgeries have a deviation greater than 50% of the real duration, whilst for the model’s forecast, only half of the surgeries do so (Figure 7, F). This relationship makes sense. Presumably, the surgeon makes forecasts considering a minimum limit for the surgery duration due to unforeseen events and complications that can occur (i.e., the surgeon does not take the risk of predicting surgeries with a duration of fewer than 30 minutes). The model does not consider a minimum limit. Further, the base unit of the two forecasts is different, which could compromise the forecast value. Most likely, the base unit for surgeon estimations is usually 5 minutes, whereas it is 1 minute or even seconds for model estimations. For example, the surgeon speculates that his surgery may take 40 or 45 minutes, but the surgeon never claims it takes 42 or 43 minutes, as the model might determine.

For outlier surgeries, we analysed short-term surgeries and long-term surgeries separately.

For short surgeries, both forecasts – model and surgeon – are consistently above the observed surgery duration in the OR for both datasets (Figure 8). Additionally, we can see in Figure 8, G that, for the January 2016 – August 2019 dataset, the shorter surgeries of the short-term surgeries – probably with durations of 1, 2 or 3 minutes – have relatively more discrepant forecasts – model and surgeon – from the rest with a deviation exceeding 10 times the real duration. These cases could be surgeries that ended up not occurring and were closed a few minutes later due to an unforeseen event right at the beginning of the surgery or surgeries with registration errors. Model and surgeon forecasts accounted for the entire intervention execution, as expected, therefore, there is a high deviation between the two durations. No deduction from the September 2019 – December 2020 dataset can be made, mainly due to the small sample size (Figure 8, H).
For long surgeries, both forecasts – model and surgeon – are practically all below the observed surgery duration in the Operating Room for both datasets, as we can see in Figure 9. In the January 2016 – August 2019 dataset, we can see that the model forecasts are almost always less than half the real surgery duration (more than 50% of error), while surgeons do not miss more than 40% of the real value (minus 40% of error) (Figure 9, I). That means the surgeon’s forecasts are considerably closer to the observed duration for long-term surgeries. The same finding is visible in the September 2019 – December 2020 dataset; very are the surgeries whose surgeon’s estimation exceeds 60% of real duration (minus 40% of error), and for the model, no forecast exceeds 40% of the real value (more than 60% of error) (Figure 9, J).

6 OPERATING ROOM SCHEDULE: MODEL IMPACT ASSESSMENT

A genetic algorithm approach, based on the Non-dominated Sorting Genetic Algorithm II (NSGA-II), to design surgery schedule was developed, trained and tested using data from HLL.

6.1 Coding of OR Scheduling with Genetic Algorithms

The proposed genetic algorithm is an introductory approach to the problem of scheduling surgery. Three hard constraints are considered when creating the solutions; soft restrictions were never considered. Hard constraints cannot be violated to prevent generating an infeasible schedule, whereas soft constraints might be broken. However, the number of violations of soft constraints must be minimised. The hard constraints considered when creating the solutions were:

1. the surgery can only be performed within a period of free time
2. the surgeon cannot perform different surgeries simultaneously
3. the patient cannot undergo more than one surgery at the same time

For the proposed algorithm, the hard constraints are used to calculate the fitness. For each hard constraint fulfilled, the score increases by one value. If a surgery violates a constraint at any time-space slot that it occupies, its score is not incremented. Thus, each surgery can have between 0 and 3 points.

The schedule is feasible if all surgeries meet the three hard constraints. Therefore, the three rules are verified for all surgeries, and the points for each surgery are concatenated with the others. In the end, a schedule is viable when the schedule score is equal to 3 times the number of surgeries allocated. Once these constraints are satisfied, the fitness function, equation (1), can be calculated. As illustrated in Figure 10, when all the surgeries meet the three hard constraints, the fitness function of the proposed genetic algorithm follows a triangular distribution.

Figure 7 - Deviation of duration compared with the observed surgery duration in the OR for each non-outlier surgery. The surgeries are sorted in ascending order according to the observed surgery duration in the OR. E: January 2016 – August 2019 dataset. F: September 2019 – December 2020 dataset. Use of the IQR method. Surgeon forecast metrics against real data are represented in the darkest colour. Luz Saíde’s model forecast metrics against real data are marked in a lighter colour. The yellow line corresponds to the 1º quartile, the orange to the 2º quartile and the red line to the 3º quartile.

Figure 8 - Deviation of duration compared with the observed surgery duration in the OR for each short-term outlier surgery. The surgeries are sorted in ascending order according to the observed surgery duration in the OR. G: January 2016 – August 2019 dataset. H: September 2019 – December 2020 dataset. Use of the IQR method. Surgeon forecast metrics against real data are represented in the darkest colour. Luz Saíde’s model forecast metrics against real data are marked in a lighter colour.

Figure 9 - Deviation of duration compared with the observed surgery duration in the OR for each long-term outlier surgery. The surgeries are sorted in ascending order according to the observed surgery duration in the OR. I: January 2016 – August 2019 dataset. J: September 2019 – December 2020 dataset. Use of the IQR method. Surgeon forecast metrics against real data are represented in the darkest colour. Luz Saíde’s model forecast metrics against real data are marked in a lighter colour.
Additionally, we verify by equation (1) and Figure 10 that the fitness value of the schedule is one when all surgeries meet the three hard constraints and there are no idle times or overtime in the OR (i.e., when the sum of the 30-minutes space slots of all allocated surgeries is equal to the total number of available slots) – ideal case.

\[
\begin{aligned}
&\text{if schedule score} \leq 3 \times \text{number of surgeries allocated} \\
&\text{Fitness function value} = 1 \quad \text{if schedule score} = 3 \times \text{number of surgeries allocated}
\end{aligned}
\]

Figure 10 - Distribution of the fitness function of the proposed genetic algorithm

The proposed genetic algorithm has five feature classes: surgeon, speciality, operating room, patient, and surgery. An ID identifies the surgeon, speciality, theatre, and patient classes. The surgery class is defined by a surgeon, speciality and patient ID, and surgery duration. Duration corresponds to the integer number of 30-minute space slots needed to perform the intervention. To account for turnover time, 15 minutes were added to the surgery duration in time units (minutes) before determining the integer number of equivalent 30-minute slots.

In the proposed algorithm, the algorithm starts by initialising a population of randomly generated chromosomes. Two-point crossover is applied in the proposed algorithm. A fitness level greater than 75% was the stopping criteria, which corresponds to 8 slots, which mean 4h of idle time or overtime beyond midnight.

6.2 Computational Experiments

Computational experiments were carried out on the cluster Sigma (a Unix Shell service provided by IT Services of the Instituto Superior Técnico).

Ideally, we intended to develop a genetic algorithm program that would weigh the priority of surgeries and design schedules, distributing the surgeries with the highest priority to the first few days. However, in the datasets provided by Luz Saúde, there is no information on the priority for performing the surgery. Therefore, this research assumed that the surgeries that would be intended to be scheduled would be indicated in advance (for example, one or two days before).

The planning of surgeries in the OR at HLL is carried out weekly. Thus, together with Luz Saúde, a scenario like an OR reality was defined for the program’s execution. It would be considered that the OR would be open for five working days, from Monday to Friday, 16 hours a day (opening hours: 08:00 – 00:00) and with 8 theatres in operation. However, the basic algorithm suffers from a high order of complexity and runtime complexity, rendering it less useful for practical applications. We decided to minimise the scenario as much as possible since this study is limited in time and computational resources. As a result, a daily schedule for an operative theatre in activity 16 hours a day would be obtained.

The output was a schedule represented as a table of one column (one working day) and 32 rows (30-minutes slot number equivalent to 16 hours). Table elements can be empty if no surgery is placed at a specific time and theatre or filled with a surgery. The surgery appears in the schedule identified by the surgical speciality, and primary surgeon and patient ID.

6.3 Identifying Efficient Genetic Algorithm Parameters

To choose the genetic algorithm parameters, a randomly chosen day from the January 2016 – August 2019 dataset was used, previously excluding the surgeries with recording errors: 16 January 2018.

Different values for population, crossover and mutation parameters were tried, and those that provided the best results were chosen. Table 3 presents the combination of selected parameters implemented in the proposed algorithm and used in all computational experiments; no changes to values were made during the algorithm’s execution (fixed values).

Table 3 - Considered values for the parameters of the proposed algorithm.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>100</td>
</tr>
<tr>
<td>Crossover prob</td>
<td>80</td>
</tr>
<tr>
<td>Mutation prob</td>
<td>2</td>
</tr>
<tr>
<td>Mutation size</td>
<td>3</td>
</tr>
</tbody>
</table>

6.4 Experimental results

The test of the proposed genetic algorithm for scheduling surgeries was performed using data of 40 days (4 weeks) from the September 2019 – December 2020 dataset, after excluding the surgeries with recording errors. The 40 days used are the working days of the periods: 30 September 2019 to 27 October 2019 and 04 November 2019 to 29 November 2019. The week 28 October to 03 November 2019 was not considered since Friday, the 1st is a national holiday, and the number of surgeries performed that week could be conditioned.

For each of the 40 days, three surgery schedule performances were created based on:
1. surgery duration forecasts from Luz Saúde’s model
2. surgery duration forecasts made by the surgeon
3. real data (i.e., observed surgery duration in the OR)
The fitness value of the schedule considering the forecasts of the surgeon and the fitness value of the schedule resulting from Luz Saúde’s model were compared with the fitness value of the schedule based on real data. On average, the schedule considering the surgeon’s forecasts and the schedule resulting from Luz Saúde’s model have a deviation of 89.69 minutes (i.e., 1 hour and 30 minutes) and 69.74 minutes (i.e., 1 hour and 10 minutes), respectively, against the schedule based on real data (as we can see in Figure 11). The results show that the model is less than 20 minutes wrong concerning the surgeon, in terms of occupation time with the surgeries. Thus, these results demonstrated that the model could optimise the scheduling.

Figure 11 - MAE (in minutes), MAPE (in parentheses) and margin of error of the deviation using 95% confidence interval of the schedule considering the surgeon’s forecast and the schedule based on Luz Saúde’s model face with the schedule resulting from real data without considering the Rooms. 40 days of September 2019 – December 2020 datasets. Surgeon forecast metrics against real data are represented in the darkest colour. Luz Saúde’s model forecast metrics against real data are marked in a lighter colour.

It should be highlighted that no outliers are between the 40 days of testing nor on 16 January 2018. As verified in section 5.1, although the forecast error of Luz Saúde’s model is relatively high for each outlier, globally, the model’s error is small since the outliers constitute less than 1% of the dataset. Therefore, due to the statistical insignificance of the outliers, it was not considered relevant to contain or not outliers on the days used for the test. Furthermore, it would be expected that better results could have been obtained since the model should be more accurate when not interacting with outliers.

For a more confident analysis, we studied deviations between schedules considering different surgery durations. For example, more complex and longer surgeries or more detailed specialities may have more significant errors once they might be more prone to complications or unforeseen events during the intervention. Consequently, their duration is more difficult to predict.

We verified that the surgeries at HLL are essentially concentrated until three hours, so a generic scenario was defined in Room X. We wanted to analyse the impact of the model’s forecast in the design of schedules for shorter and longer surgeries. A scenario with a higher proportion of surgeries up to one hour was defined in Room Y and a scenario essentially with long surgeries was defined in Room Z. Table 4 summarises the surgery characteristics considered in each of Rooms X, Y and Z. For each of the 40 days, three operating theatres were chosen at random but restricted to the specific surgery characteristics described above and in Table 4. Then, we investigate the dominant specialties in each room, data also presents in Table 4.

We confront the fitness value of the schedule considering the forecasts of the surgeon against the fitness value of the schedule based on real data. The exact process was carried out for Luz Saúde’s model. As expected, the results were different for each room, as shown in Figure 12. They will be analysed below.

Table 4 - Dominant characteristics and specialties in each room considered to study the impact of Luz Saúde’s model on the design of surgery schedules.

<table>
<thead>
<tr>
<th>Surgery</th>
<th>Specialty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Room X</td>
<td>Surgeries lasting more than 3 hours are less than 25% of all surgeries. General surgery orthopaedics, ophthalmology</td>
</tr>
<tr>
<td>Room Y</td>
<td>Surgeries lasting up to 1-hour account for more than 45% of all surgeries, and surgeries lasting more than 3 hours do not exceed 15%. Ophthalmology and orthopaedics</td>
</tr>
<tr>
<td>Room Z</td>
<td>Surgeries lasting more than 3 hours are greater than 65% of all surgeries. Orthopaedics, general surgery and urology</td>
</tr>
</tbody>
</table>

We started by performing a comparison of deviations between the rooms. As it is visible in Figure 12, the percentage errors for long surgeries (Room Z) are superior. Between the surgeon and real data (dark blue bars), there is a deviation of 15.47% for Room Z against 6% for Room X and Room Y. Between the model and real data (light blue bars), there is a deviation of 9.62% for Room Z against 7%/6% for Room X and Room Y. This result is predictable because, as visible in Figure 5, for the September 2019 – December 2020 dataset, the model and surgeon forecasts have a higher MAE and MAPE for longer surgeries which will have a cumulative impact on the design of schedules.

Figure 12 - MAE (in minutes), MAPE (in parentheses) and margin of error of the deviation using 95% confidence interval of the schedule considering the surgeon’s forecast and the schedule based on Luz Saúde’s model face with the schedule resulting from real data for each of the Rooms X, Y and Z. 40 days of September 2019 – December 2020 datasets. Surgeon forecast metrics against real data are represented in the darkest colour. Luz Saúde’s model forecast metrics against real data are marked in a lighter colour.

We also performed an analysis per room.

As we can see in Figure 12, for Room X, the deviation between the schedule considering the forecasts of the surgeon and the schedule based on real data (dark blue bar) is 57.59 minutes. The deviation between applying the model and using real data (light blue bar) is 63.45 minutes. Comparing the values, the model’s schedule differs 6 minutes from the schedule resulting from the
surgeon forecast. For Room Y, the deviation between considering the surgeon’s forecasts and real data (dark blue bar) is 62.97 minutes. The deviation between applying the model and using real data (light blue bar) is 53.22 minutes (Figure 12). Comparing the two values, the model’s schedule differs only in 10 minutes from the schedule resulting from the surgeon forecast. For Room Z, the deviation between considering the surgeon’s forecasts and real data (dark blue bar) is 148.5 minutes. The deviation between applying the model and using real data (light blue bar) is 92.31 minutes (Figure 12). Comparing the two values, the model’s schedule differs 56 minutes from the schedule resulting from the surgeon forecast.

From the previous analyses, we could conclude a more considerable discrepancy between the schedules – model versus surgeon – for long surgeries (56 minutes for Room Z versus 6 minutes and 10 minutes for Rooms X and Y, respectively). Furthermore, for each room, we drew the fitness function as a function of the sum of the slots of all surgeries allocated in the schedule (Figure 13, Figure 14 and Figure 15) and marked the three schedules under study (considering the model and surgeon forecasts and the real duration). As aforementioned, the fitness function of the genetic algorithm follows a triangular distribution (Figure 10).

As we can see in Figure 13, the forecasts made by the model and those performed by the surgeon were by excess concerning the real surgery duration in the OR for Room X. On the other hand, we can see in Figure 14 that the forecasts made by the model and those performed by the surgeon were by default concerning the real surgery duration for Room Y. These observations are consistent with those shown in Figure 6. For non-outlier surgeries (lasting between 12 minutes and 8 hours), there is no direct relationship in terms of overestimation or underestimation for the model and surgeon forecasts.

Additionally, we studied in Figure 7 that the model’s forecasts are closer to the real duration than the surgeon’s forecasts for non-outlier surgeries. That is also visible in Figure 14. It is also evident in Figure 13 and Figure 14 that the model and surgeon schedules are close in terms of occupation time with the surgeries for the two rooms. That is consistent with the mentioned proximities of 6 and 10 minutes for Rooms X and Y, respectively.

For Room Z, both forecasts – model and surgeon – are by default concerning the real duration of the surgery in the OR, as we can see in Figure 15. They are consistent with the study for long surgeries in Figure 9. Plus, in Figure 9 we visualised that the surgeon’s forecasts are considerably closer to the real duration than the model’s forecast, which is also evident in Figure 15. Illustrated in Figure 10, for each of those days, the fitness value of the schedule based on real data is on the right side of the triangular (for values greater than 32).

Figure 13 - Fitness function of the proposed genetic algorithm as a function of the sum of slots of all surgeries allocated in the schedules for Room X.

Figure 14 - Fitness function of the proposed genetic algorithm as a function of the sum of slots of all surgeries allocated in the schedules for Room Y (essentially short surgeries up to one hour).

Figure 15 - Fitness function of the proposed genetic algorithm as a function of the sum of slots of all surgeries allocated in the schedules for Room Z (essentially long surgeries).

Finally, it is needed to make an additional highlight to Room Z. Unlike Room X and Room Z, a significant proportion of the 40 days has the OR in activity beyond midnight for Room Y. That mean, for each of those days, the sum of slots for the real duration of the allocated surgeries was greater than 32 (the total number of slots available until midnight). Since the fitness function of the proposed genetic algorithm follows a triangular distribution (as On the other hand, the model and surgeon forecasts are by default concerning the real duration for long surgeries. Hence, the fitness values of the schedules considering the forecasts are found on the left side of the triangular (for values less than 32), as we also can see in Figure 15.

Then it should be noted that although the schedule resulting from the model has a lower deviation compared with the schedule considering the surgeon’s forecasts for Room Y, which could not mean that the model schedule is more trustworthy with reality. The explanation for this fact lies in the triangular distribution that follows the fitness function. As shown in Figure 10, a slight deviation can mean two points with close ordinate and close abscissa (the two abscissas on the same side of the triangle) or
two points with close ordinate and distant abscissa (the two abscissas on different sides of the triangle). Thus, and as we can see in Figure 15, the surgeon’s forecast is the closest to the real duration (closer abscissas but more distant ordinate). However, it is the model’s forecast that corresponds to a fitness value closer to the fitness value of the schedule based on real data (more distant abscissas but closer ordinate). Consequently, it will be the schedule resulting from the model that will have the slightest deviation. Given this point, the selected fitness function, equation (1), may not have been the best bet.

7 FINAL CONSIDERATIONS OF THE CASE STUDY, LIMITATIONS AND FUTURE WORK

Several benefits are identified with the application of the Luz Saúde’s model. In addition to being a self-sustaining and fast process, it takes advantage of unlimited amounts of multidimensional and multivariate data at once. Furthermore, the model evolves in quality and efficiency as it stores newer data (while the surgeon will hardly update his forecasts). The opportunity to reduce the time used to execute the scheduling task is also an outstanding advantage identified by Luz Saúde. From the perspective of this research, it is perfectly feasible to replace the current method for predicting the surgery duration with Luz Saúde’s model. The risk for the HLL of using the results of the current model or the model is practically the same.

Moreover, in this research, genetic algorithms are also recognised as an appreciable tool for designing the OR schedules. However, more effort in future research will still be required to optimise the genetic algorithms and overcome the high-order complexity problem. To be applicable, it will also require the introduction of new restrictions (e.g., surgeon A only operates on day Y). However, it is important to mention that there were some limiting factors across the development of this research, namely the limited access to real data, the computational complexity characteristic of genetic algorithms, and limited time and computational resources. If this research were carried out in a more resourceful computation environment and using more or improved (cleaner) data, we would possibly have achieved better results and perhaps a shrewder model.

Despite the results, it is important to highlight that there is a diversity of domain expertise still to be explored in future research, which could contribute to the model’s improvement in the certainty and confidence of its forecasts. More assertive forecasts may also be obtained in the future with the application of other training conditions and strategies of feature generation.

8 CONCLUSIONS

From the results obtained it was verified that it is perfectly feasible to adopt Luz Saúde’s model to predict the surgery duration. Also, the results of the model deployment in the planning of elective surgeries have demonstrated that the model could optimise scheduling. Although the proposed algorithm addresses the surgery scheduling problem in a very simplistic way, genetic algorithms have shown to be a promising tool for designing the surgery schedules. Overall, this research illustrates a remarkable advancement in applying ML techniques in decisive operational problems in medicine. Using the model along with the genetic algorithm is an innovative approach for OR planning.

9 PREFACE

This research was performed at the GLSMED Learning Health, S.A. (Lisbon, Portugal), during the period March – October 2021, under the supervision of Professor Cláudia Antunes (Instituto Superior Técnico). The thesis was co-supervised by Ana Bento (specialist of Strategy and Business Analytics at Luz Saúde).

10 REFERENCES