

# INTRAOPERATIVE REMAINING SURGERY DURATION ESTIMATION TO IMPROVE OPERATING ROOM SCHEDULING

---

K.N.M.M.H. Osman





# Intraoperative remaining surgery duration estimation to improve operating room scheduling

*The creation and evaluation of an estimation system*

by

Karim Nehad Mohamed Mohamed  
Hamed Osman

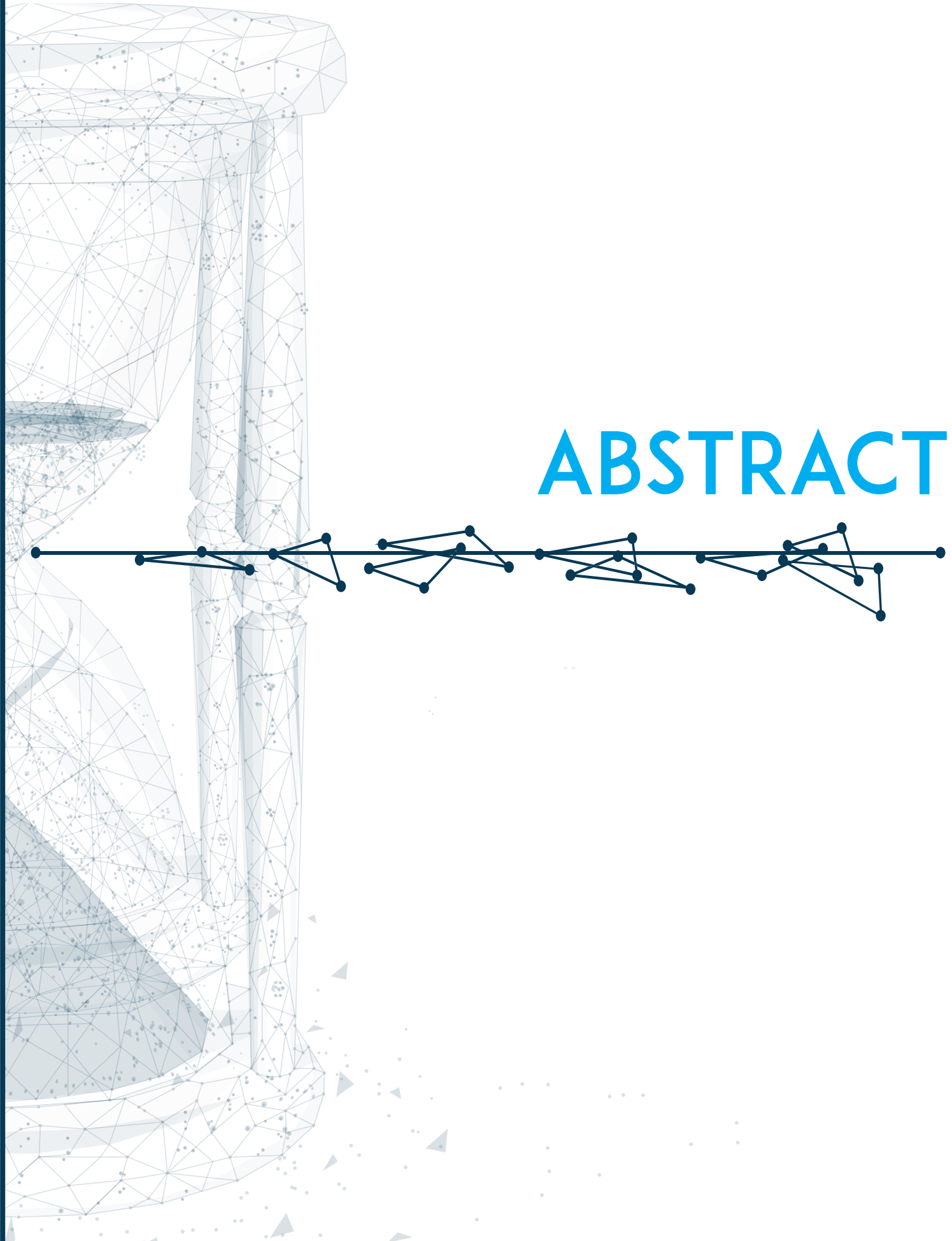
to obtain the degree of Master of Science  
at the Delft University of Technology,  
to be defended publicly on Thursday the 16<sup>th</sup> of July, 2020 at 10:00 AM.

|                   |                            |                               |
|-------------------|----------------------------|-------------------------------|
| Student number:   | 4309014                    |                               |
| Project duration: | January, 2020 – July, 2020 |                               |
| Thesis committee: | Dr. J.J. van den Dobbelen  | TU Delft (supervisor)         |
|                   | Prof. dr. B.H.W. Hendriks  | TU Delft                      |
|                   | Dr. ir. M.C. Goorden       | TU Delft                      |
|                   | Dr. ir. A.C.P. Guédon      | Spaarne Gasthuis (supervisor) |

*This thesis is confidential and cannot be made public until July 16, 2022*

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

# ABSTRACT



One of the main elements for creating an optimal operating room schedule is an accurate surgery duration estimation. Currently, this estimation is only done preoperatively. However, multiple factors during the surgery itself could influence the duration, for example bleeding. Literature showed the possibility of intraoperatively estimating a surgery duration based on surgical progress. However, one of the main concerns was the usability of such a system in the operating room workflow. Therefore, this research was focused on two parts: (1) the creation of an automatic intraoperative remaining surgery duration estimation system and (2) the evaluation of such a system for the operating room workflow. Two types of surgeries were used for the estimation, the Laparoscopic Cholecystectomy and the Total Laparoscopic Hysterectomy. The estimation was created using multiple statistical regressor methods, such as linear regression and random forest, and progress-based methods based on the nearest-neighbors algorithm and Dynamic Time Warping method. The evaluation was done on two levels: the system level based on the error of the estimation, and the operating room workflow based on surgical data from 2016 to 2019 and interviews with the operating room program coordinators. Results showed that an intraoperative remaining surgery duration estimation system based on surgical phases was able to re-estimate the duration with an error of about 10 minutes, an acceptable error for the operating room workflow. Moreover, the third quarter of the surgery showed to be the essential part where an accurate estimation is needed. Furthermore, an automatic system showed additional benefits such as being unbiased, continuous, and reducing unnecessary disturbance in the operating rooms. Overall, this research showed that an intraoperative remaining surgery duration system based on surgical phases is promising for the operating room workflow. Future research is needed to understand how to implement such a system in the operating room workflow.

The background features a complex geometric design. On the left, a large, semi-transparent wireframe structure resembling a classical column or architectural element is visible. A network graph, consisting of black nodes and connecting lines, extends horizontally across the middle of the page. The word 'PREFACE' is printed in a bold, blue, sans-serif font on the right side. The overall aesthetic is clean, technical, and modern.

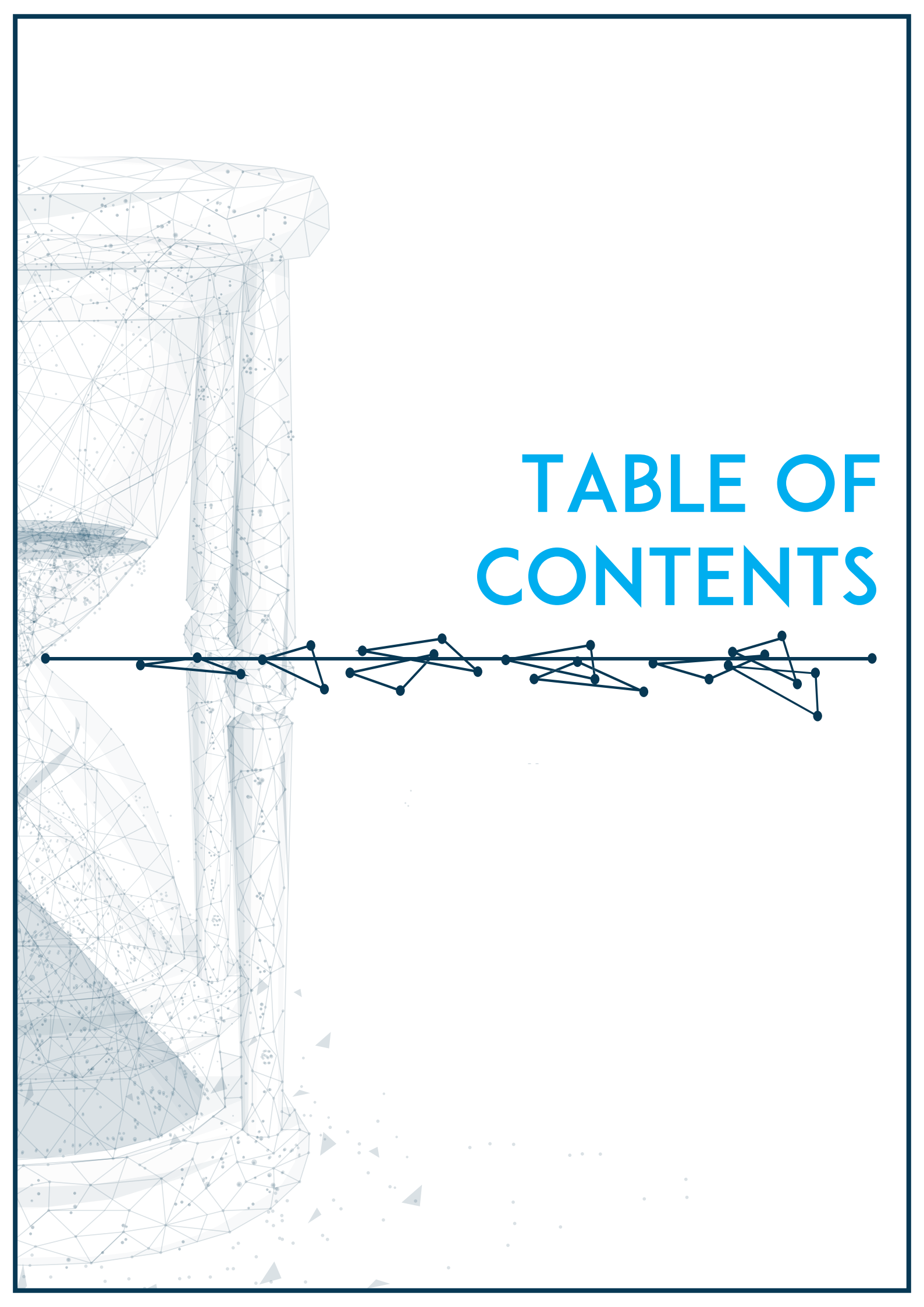
# PREFACE

To stay in terms of this research, at the beginning I estimated to have about six months as the total duration for my thesis. I scheduled each part correspondingly in the following order; a literature study phase, method design phase, creation phase, and evaluation phase. However, a large complication occurred in the world that changed the procedure of this research; COVID-19. This happened at the end of the creation phase, with the consequence that I was not allowed back in the hospital. Evaluating the operating room workflow without being able to observe the workflow created challenges, challenges that due to the support of Annetje and the kind Spaarne employees I was able to successfully overcome. By following a different method and research procedure, I was still able to finish this research in the estimated thesis duration. Without the data from Business Intelligence and the interviews with the operating room program coordinators, the evaluation of the operating room workflow would have never been possible, which I am deeply thankful for.

Most importantly, I sincerely would like to thank dr. ir. Annetje Guedon, as she was my daily supervisor and supported me every step of this research. I am really grateful for the many meetings and phone calls, and all the helpful feedback I received. Furthermore, I would like to thank dr. John van den Dobbelsteen as being my university supervisor, first of all for recommending this research with Annetje, but also for the feedback and guidance through the entire process. Next, I would like to thank dr. Andru Twinanda and Helena Kloosterman from COSMONiO for their guidance with the method design and the feedback on the system. Also, I would like to thank prof. dr. Benno Hendriks and dr. ir. Marlies Goorden for being committee members and taking the time to understand my thesis and giving me the possibility to defend it.

Lastly and possibly the most important people, I would like to thank my parents. Since the beginning of elementary school, many teachers told them that I was not smart enough. However, that did not stop them to challenge me to reach my full potential, potential they saw before anyone else. As they just moved from Egypt, the language and culture difference made it especially hard for them to educate me but also themselves properly. However, they always pushed me to learn and understand the world to better myself, and with all the freedom I needed to think for myself. I think that my parents are even more excited about me getting my degree than myself, and I am grateful that I am able to give them this joy. Thank you, both of you.

*K.N.M.M.H. Osman  
Delft, July 2020*

The background features a complex geometric design. On the left, a vertical wireframe structure resembling a classical column is rendered in a light blue, semi-transparent style. To the right, a network diagram consists of a horizontal line with several nodes, from which various geometric shapes like triangles and polygons branch out. The overall aesthetic is clean, technical, and modern.

# TABLE OF CONTENTS



|   |           |
|---|-----------|
| <b>Abstract</b>                           | <b>4</b>  |
| <b>Preface</b>                            | <b>6</b>  |
| <b>Table of contents</b>                  | <b>8</b>  |
| <b>Introduction</b>                       | <b>10</b> |
| <b>Method</b>                             | <b>14</b> |
| 2.1. Dataset                              | 15        |
| 2.1.1. Smoothing phases                   | 17        |
| 2.2. System architecture                  | 18        |
| 2.3. Current estimation methods           | 18        |
| 2.3.1. Naive approach                     | 18        |
| 2.3.2. Phase-Inferred method              | 19        |
| 2.3.3. Linear Regression                  | 19        |
| 2.3.4. Multilayer Perceptron Regression   | 20        |
| 2.3.5. Decision Tree Regression           | 21        |
| 2.3.6. Random Forest Regression           | 22        |
| 2.4. Novel method                         | 22        |
| 2.5. Evaluation                           | 24        |
| 2.5.1. System level evaluation            | 24        |
| 2.5.2. Operating room workflow evaluation | 25        |
| <b>Results</b>                            | <b>28</b> |
| 3.1. System level                         | 29        |
| 3.2. Operating room workflow level        | 42        |
| 3.2.1. Turnover time                      | 42        |
| 3.2.2. Overtime                           | 44        |
| 3.2.3. Undertime                          | 44        |
| 3.2.4. Interview                          | 47        |
| <b>Discussion</b>                         | <b>50</b> |
| 4.1. System evaluation                    | 51        |
| 4.2. The current OR workflow              | 53        |
| 4.3. Implementation in the OR workflow    | 55        |
| 4.4. Future research                      | 55        |
| 4.5. Conclusion                           | 56        |
| <b>Bibliography</b>                       | <b>58</b> |
| <b>Appendix A - Interview results</b>     | <b>62</b> |
| Interview 1                               | 63        |
| Interview 2                               | 65        |
| Interview 3                               | 67        |



# INTRODUCTION

Hospitals have many disciplines working together for one main goal: healing patients. Each department has its protocols, and they work almost independently of each other. However, there is one place where many disciplines meet, the operating room. The operating rooms are used by many departments, where each department has specialized types of surgeries. This varies from microscopic surgery in Neurology, to hip surgery in Orthopaedics. To be able to work together with all departments within the operating rooms, hospitals work with an operating room schedule. The operating room schedule indicates which surgeon needs to perform what type of surgery on which patient. It indicates where to perform the surgery, but also the time to start. It is the thread that links everything together. Therefore, an optimal operating room schedule is of great importance.

As all scheduling problems [1–4], the duration of a task is the main factor that influences the creation of a schedule. As some problems can restrict the duration of tasks, such as school scheduling [5] or the classical traveling-salesperson problem [6], surgeries always need to finish with no restrictions on time. The planning of the operating room should not influence the surgeon and the staff decision making regarding healthcare. They should have the freedom to operate in the time and speed they deem necessary for the care of the patient. So the following problem arises; how to create an operating room schedule without predefining the duration of the surgeries. This is done by a surgery duration estimation.

Currently, hospitals estimate the surgery duration preoperatively. This is done with different methods, usually using historical surgical times to estimate the duration [7] or sometimes patient characteristic based estimation [8–10]. Based on an analysis of the Spaarne Gasthuis hospital in the Netherlands, as can be seen in Table 1.1, the real surgery duration still differs significantly compared to the preoperative estimated duration. Surgeries deviate from the preoperative estimated surgery duration with on overall a standard deviation of 36 minutes. This shows that using preoperative surgery duration estimation is challenging, as one of the reasons may be that the duration changes during the course of the surgery itself. The speed of the surgeon could be a factor, but also events occurring in the surgery such as bleeding or a change in protocol due to new findings.

The use of intraoperative surgery duration estimation could give a better indication of the progress of the surgery and eventually result in a better operating room schedule. Intraoperative estimation is currently done manually. The operating room staff asks the surgeon to estimate the surgery duration during the surgery. However, this has disadvantages. First, surgeons could be biased [11], as the estimation is only based on the opinion of the surgeon. Another issue is that manual estimation is based on one moment. The operating room staff retrieves the estimation from a surgeon at a specific time by entering or calling the operating room. This could be a moment just before an event, such as bleeding, that influences the duration significantly. Furthermore, entering or calling the operating room for an estimation causes unnecessary interruptions that could have negative consequences on the surgery [12, 13]. These disadvantages of manual intraoperative estimation could be solved by using an automatic intraoperative estimation system.

Research showed that progress based systems could improve the estimation of the remaining surgery duration [14–17]. This progress can be described in phases, where each phase is a

Table 1.1: Descriptive statistics of the surgeries at the Spaarne Gasthuis hospital from 2016 to 2019. It contains the number of surgeries, actual surgery durations and the error deviation for the estimated surgery duration using current estimation calculation, grouped by specialism. A positive number represents a delayed surgery and a negative represents less time needed for a surgery

|                                | N     | Total duration |       | Estimation error |       |
|--------------------------------|-------|----------------|-------|------------------|-------|
|                                |       | Mean           | Std   | Mean             | Std   |
| Overall                        | 56396 | 59.85          | 44.51 | 3.73             | 36.23 |
| General Surgery                | 22745 | 66.46          | 48.45 | 6.96             | 42.56 |
| Gynaecology                    | 5730  | 47.17          | 32.26 | 1.04             | 28    |
| Otorhinolaryngology            | 3979  | 37.9           | 39.95 | -0.98            | 25.08 |
| Oral and Maxillofacial Surgery | 1629  | 58.33          | 46.3  | -3.14            | 35.37 |
| Neurosurgery                   | 1634  | 61             | 33.21 | 2.86             | 30.83 |
| Ophthalmology                  | 303   | 37             | 19.11 | -4.02            | 23.52 |
| Orthopaedic Surgery            | 11074 | 59.08          | 34.24 | 2.91             | 23.78 |
| Plastic Surgery                | 4576  | 70.03          | 51.48 | 0.14             | 42.23 |
| Urology                        | 4641  | 56.09          | 48.18 | 4.1              | 39.07 |

specific part of the surgical procedure, such as the "closing" phase. However, a problem with an automatic system is that it needs the current phase of the surgery as input. Some research was done in using tool detection to predict the phase or progress, such as using the activation pattern of electrosurgical devices [18], or operating room sensor data [19]. However, using such recognition data alteration or addition of systems in the operating room or in the tools are needed to be able to detect these signals.

Meij [20] created in collaboration with COSMONiO a recognition network that uses endoscopic video to predict the tools used in the video, which is based on the RSDnet neural network of Twinanda et al. [17]. Using these tools as input, the network could predict the phase of each frame in a video. However, it was not known how these predicted phases could be used to estimate a remaining surgery duration. Furthermore, questions were raised concerning how to evaluate such a system, on a system level, but also on a more practical level for the operating room workflow. For example, what kind of accuracy is needed so that a system would be useful for the operating room workflow? For that reason this graduation thesis focussed on two parts:

1. The *creation* of an intraoperative surgery duration estimation system based on surgical phases.
2. The *evaluation* of such a system for the use in the operation room workflow.

The goal of this thesis was to understand if an automatic intraoperative surgery duration estimation system could be created using predicted phases from the recognition network and to evaluate the benefit of such a system for the operating room workflow. For that reason, the system was created as explained in chapter 2 with existing methods (section 2.3) and a novel method for this field (section 2.4) to discover the possibility of developing such a system. Furthermore, the system was evaluated as described in section 2.5 on a systematic level (subsection 2.5.1) and the potential added value of the system in the operating workflow (subsection 2.5.2) based on a data analysis and expert interviews. The results of these methods, the operating room data analysis and the interviews are shown in chapter 3. Chapter 4 discusses these results and the potential addition of an automatic intraoperative remaining surgery duration estimation system for the operating room workflow.



# METHOD



The following chapter explains the methods that were used for this research concerning the estimation of the remaining surgery duration and the evaluation of such a system. Section 2.1 explains the datasets that were used for this research, and the properties of these datasets. Section 2.2 explains the architecture of the system. Section 2.3 defines the current statistical methods that were used and provides an explanation of these methods. Section 2.4 describes a new methodology used for the estimation of the remaining surgery duration, which is based on multiple statistical approaches. Section 2.5 describes how the methods were evaluated on a system level, and the evaluation of the operating room workflow.

## 2.1. Dataset

The data that was used in this thesis was partly retrieved from the previous research of Meij [20] and further gathered during this thesis. The data consisted of videos with two types of surgeries: the Total Laparoscopic Hysterectomy and the Laparoscopic Cholecystectomy procedure. The videos of these surgeries were retrieved from the database of the Spaarne Gasthuis hospital in the Netherlands. The set consists 33 Laparoscopic Cholecystectomy and 36 Total Laparoscopic Hysterectomy videos, with a variety of five and three different surgeons respectively. The data also varies in surgery duration (Figure 2.1). Each video consists of manual annotated phases and tools, and predicted phases based on the neural network as described by Meij [20]. The phase prediction network was a convolution neural network created by COSMONiO, which is an adaptation of the network created by Twinanda et al. [17]. Each video was split into frames, starting from first incision to closing. A frame is a moment in the video. Normally a video consists of about 20 to 60 frames per second, but for this research one frame is used for one second. This reduction of frames was necessary due to the otherwise large amount of data and computation time. This was done however without losing a significant amount of information. The phase detection neural network predicted the phase of the current frame. The phases for the Laparoscopic Cholecystectomy and Total Laparoscopic Hysterectomy are described in Table 2.1 and Table 2.2 respectively. No phase can be performed simultaneously, except the bleeding phase. The phases should occur in sequential order, however, it is possible to have surgeries where a surgeon returns to a previous phase.

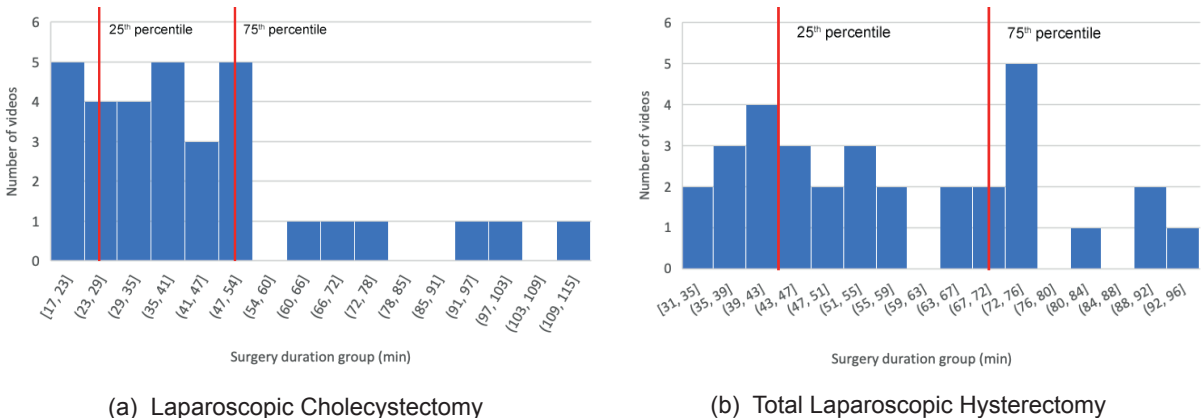


Figure 2.1: Frequency table for the Laparoscopic Cholecystectomy and Total Laparoscopic Hysterectomy videos based on surgery duration indicating the 25<sup>th</sup> and 75<sup>th</sup> percentile

Each frame consisted of the properties as described in Table 2.3. The properties were based on information from the video data and the output of the phase detection neural network. The performance of the network was assessed in terms of recall and precision. The average recall and precision were weighted in respect to the number of frames of the specific phase. It had a weighted recall of 0.77 and weighted precision of 0.79 for the phase detection of the Laparoscopic Cholecystectomy, and a weighted recall of 0.79 and weighted precision of 0.78 for the Total Laparoscopic Hysterectomy.

Table 2.1: Surgical phases for the Laparoscopic Cholecystectomy procedure

| Phase                                | Start cue  | End cue  |
|--------------------------------------|--|--|
| 1. Trocar & tools insertion          | First frame with a view of the inside of the body                                | First frame with a tool in view                            |
| 2. Preparation & dissection          | Frame after first tool in view   | Frame before the clipper is in view                        |
| 3. Clipping & cutting                | First frame with the clipper in view   | Last frame with the scissors in view                       |
| 4. Gallbladder dissection            | Frame after the last scissors in view  | Frame before the bag is in view                            |
| 5. Gallbladder packaging & retrieval | First frame with the bag in view   | Last frame with the bag in view                            |
| 6. Liver bed coagulation             | First frame where the grasper is used to coagulate                               | Last frame the grasper is used to coagulate                |
| 7. Final check & irrigation          | Frame after the last frame of coagulating or the last frame with the bag in view | Last frame before removing the trocars or leaving the body |
| 8. Closing & desufflation            | First frame of removing the trocars or leaving the body                          | First frame outside the body or the end of the video       |
| <i>Additional: Bleeding</i>          | First frame with blood and the irrigator or gauze                                | Last frame with blood and the irrigator or gauze           |

Table 2.2: Surgical phases for the Total Laparoscopic Hysterectomy procedure

| Phase                                | Start cue   | End cue  |
|--------------------------------------|---|--|
| 1. Trocar & tools insertion          | First frame with a view of the inside of the body                         | First frame with a tool in view                                    |
| 2. Uterus dissection                 | Frame after first tool in view  | Frame before the hook is used on the vaginal cuff                  |
| 3. Uterus separation from the vagina | First frame the hook is used on the vaginal cuff                          | First frame the uterus is fully separated from the vagina          |
| 4. Uterus retrieval: transvaginal    | First frame after the uterus is fully separated from the vagina           | Last frame with the bag in view                                    |
| 5. Uterus retrieval : morcellation   | First frame after the uterus is fully separated from the vagina           | Last frame with the bag in view                                    |
| 6. Vaginal cuff closure              | First frame the uterus and/or bag is not in view                          | Last frame the needle feeder and/or needle with thread are in view |
| 7. Final check & irrigation          | First frame after the needle feeder and/or needle with thread are in view | Last frame before removing the trocars or leaving the body         |
| 8. Closing & desufflation            | First frame of removing the trocars or leaving the body                   | First frame outside the body or the end of the video               |
| <i>Additional: Bleeding</i>          | First frame with blood and the irrigator or gauze                         | Last frame with blood and the irrigator or gauze                   |

Table 2.3: Video frame properties

| Frame property         | Description  |
|------------------------|--|
| Frame number           | The index of the frame, which is also the total elapsed duration |
| Phase                  | The phase that the frame is currently in                         |
| Elapsed phase duration | The elapsed duration in the specific phase                       |
| Bleeding               | If there is bleeding in that specific frame or not               |



### 2.1.1. Smoothing phases

The phase was the main property used for the prediction. The neural network predicted a probability for each phase based on the data of the frame, where the phase with the highest probability was used as the label for the frame. However, some frames were hard to classify to a specific phase, such as a frame where a tool is out of view for a second. To be able to classify the frame correctly, smoothing could be applied to use previous predictions to estimate the current phase. As shown in Figure 2.2 (green is the ground truth phase, orange the predicted phase), the addition of smoothing reduced wrongly classified jumps. It was based on the probabilities and the mode of the previous  $N$  frames, information that would also be available in a real-time setting. The pseudocode is described in algorithm 1. For each frame, the probability of the phase prediction was evaluated. If it was higher than the predefined 0.8, the network was confident enough about the phase. Otherwise, it would take the mode of the last window of size  $N$  from the predictions. The use of 0.8 as the confidence threshold and  $N$  number of frames for the mode, was based on cross validation of multiple settings. The  $N$  was defined for the Laparoscopic Cholecystectomy on 15 and for the Total Laparoscopic Hysterectomy on 27. As seen in Figure 2.2, a double smoothing was used for this dataset. The second smoothing was used to remove outlier peaks that were still in the dataset. The  $N$  used for the second smoothing was set for the Laparoscopic Cholecystectomy on 11 and for the Total Laparoscopic Hysterectomy on 15.

---

**Algorithm 1:** Phase smoothing pseudocode

---

```
for (probability, prediction) in network_output do  
  if probability < 0.8 then  
    prediction = mode(previous  $N$  predictions)  
  else  
    prediction = prediction  
  end if  
  predictionlist.append(prediction)  
end for  
return Smooth predictionlist with mode
```

---



Figure 2.2: Example of smoothing of phases as created by COSMONiO, with the time-step on the x-axis and phase on the y-axis. The green line is the ground truth phase and the orange line is the predicted phase. This figure depicts a double smoothing, with the first  $N$  for the window set on 15 and the second on 11.

## 2.2. System architecture

The remaining surgery duration estimation system was created in the Python programming language [21]. The main packages used were the NumPy and Pandas package for data retrieval and manipulations, and the open source Scikit-learn package [22] for the regression analysis methods as described in section 2.3. Furthermore, the code was run on a MacBook Pro 2013 with a 2,6 GHz Dual-Core Intel Core i5 processor. Also, the computation time an estimation for one frame was restricted to a maximum of one second.

## 2.3. Current estimation methods

Existing methods were implemented to be able to compare the performance of the estimation. The methods are described in Table 2.4 and in more detail in the following subsections. All methods were implemented with benchmarking as main intention. Settings, if possible, were optimised to find the best results for each method.

Table 2.4: Remaining surgery duration estimation methods with for each the input, description and reference to literature where the method was used before

| Method                            | Input                                    | Description   | Literature   |
|-----------------------------------|--|---|--------------|
| Naive approach                    | Preoperative surgery duration estimation | Estimation based on predefined total duration.                      | [15–17, 23]  |
| Phase inferred method             | Phase and elapsed duration               | Estimation based on statistical evaluation of all videos and phase. | [17]         |
| Linear Regression                 | All frame properties (Table 2.3)         | Frame based estimation based on trained weights.                    | [15, 16, 23] |
| Multi Level Perceptron Regression | All frame properties (Table 2.3)         | Frame based estimation using perceptron layers.                     | [16]         |
| Decision Tree Regressor           | All frame properties (Table 2.3)         | Frame based estimation based on multiple conditions.                | [16]         |
| Random Forrest Regression         | All frame properties (Table 2.3)         | Average of multiple decision trees with random start conditions.    | [16]         |

### 2.3.1. Naive approach

The naive approach is based on the predefined surgery duration, which is estimated preoperatively. The remaining surgery duration (RSD) was calculated as following:

$$RSD = \max(0, t_p - t_{el}) \quad (2.1)$$

where  $t_{el}$  is the current elapsed time, and  $t_p$  is the predefined total duration of the surgery. It is limited to zero, as the remaining duration can never be negative. The naive approach is the basic method used in most systems currently in practice. An estimation is made preoperatively, and is not changed during the duration of the surgery. Each second the elapsed time is subtracted from the predefined surgery duration. This method could only perform well if the preoperative duration estimation is optimal. However, events occurring during the surgery, which could influence the surgery duration significantly, are not taken into account with the

naive approach. The naive approach is commonly used as a baseline to evaluate the performance of an estimation system, as the data is always available for each surgery and simple to calculate. For this research, the preoperative estimated duration was retrieved from the Spaarne Gasthuis hospital database and was based on five past surgeries of the same type and surgeon [9].

### 2.3.2. Phase-Inferred method

The phase-inferred method, as described by Twinanda et al. [17], uses the information available in the current frame to estimate the remaining surgery duration (RSD), where:

$$RSD = \max(0, (t_{ref}^p - t_{el}^p)) + \sum_{m=p+1}^N t_{ref}^m \quad (2.2)$$

with  $p$  defining the current phase of the frame,  $t_{ref}^p$  the mean or median duration for the current phase  $p$ ,  $t_{el}^p$  the elapsed time in the current phase  $p$ , and the sum of the mean or median durations of all other phases  $t_{ref}^m$ . The phase-inferred method makes use of other surgeries and their phases to estimate the remaining surgery. The mean or median is calculated based on the durations of the phases of the training set. These durations are then used as reference for the remaining surgery duration. The calculation re-evaluates the progress, so that if for example phase one is finished, it only uses the durations of phase two until the end. The right choice of using the mean or median differs for the type of dataset, where the mean represents an equal average for all videos, and the median disregards or filters outliers. This comes with the cost that some outliers will be unused which could cause a loss of information. An advantage of the phase-inferred method is that it uses intraoperative information to estimate the remaining surgery duration. When the phase changes it will automatically change the estimation. However, the disadvantage is that the method needs to know the current phase of the surgery and the annotations of all the previous surgeries. For this research, this information is available (see section 2.1). Another disadvantage is that this method uses the statistical description of all surgeries of the same type, which will result in using the same average for each phase for all the estimations. This phase-based method has been researched in the past, as for example Franke et al. [14], who showed promising results. The main issue was the need to manually recognize the current phase, which had been solved due to the phase prediction network.

### 2.3.3. Linear Regression

A commonly used statistical approach for regression problems is linear regression [24]. Linear regression tries to determine constant weights for each predictor to be able to find the nearest optimal. Basically, linear functions are created for each predictor to determine the positive or negative influence on the target. In this case this was the remaining surgery duration (RSD). The linear regression model will create a function, such as:

$$RSD = c + ax_1 + bx_2 + \dots \quad (2.3)$$

where  $c$  is the intercept (constant) variable,  $a$  the weight for predictor  $x_1$ , and  $b$  the weight for predictor  $x_2$ . In this research, the predictors for the linear regression were the frame properties as described in Table 2.3. One of the main advantages of linear regression is the ability to use weighted linear correlations between features (or predictors). Also, the weights clearly define the influence of a predictor, so in practice it is a simple method to explain which predictors are influencing the remaining surgery duration positively or negatively. However, this is only useful if there is a linear correlation, since the final result will only increase or decrease in a linear structure [25], as can be seen in Equation 2.3. Furthermore, the method is only based on the

features of one frame. This causes the method to disregard previous frames and only calculate a remaining surgery duration based on the current frame. The frame properties did include the frame number and elapsed duration which gave an indication of the progress. However, the method does not use the results of previous predictions. For the implementation of the linear regression model this research used the *sklearn.linear\_model* package, which is part of the Sci-kit package.

### 2.3.4. Multilayer Perceptron Regression

The basic idea of the multilayer perceptron regression is the use of neurons that can be activated if the predefined activation threshold is achieved [27]. Each neuron has a specific weight for each input and combined with these inputs, the activation function results in either a one or a zero, depending if the activation threshold was achieved. The neurons are aligned in a layered network, where each layer is input for the next layer. If there would be only an input and output layer, the multilayer perceptron regression model functions the same as a linear regression model. However, a multilayer perceptron regression model can have multiple hidden layers, which creates a more complex network as can be seen in Figure 2.3. For the estimation of the

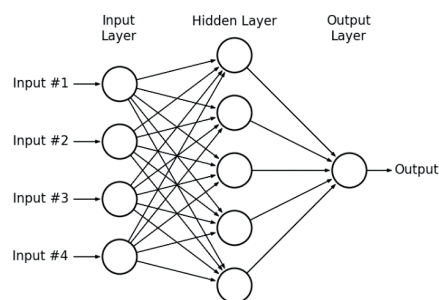


Figure 2.3: A hypothetical example of a Multilayered perceptron network [26]

remaining surgery duration, the multilayer perceptron regression model was trained on the training set. By the use of gradient descent [28], the weights for each neuron were calculated. A disadvantage of multilayer perceptron regression models, the same as the linear regression model, is that they are not able to make use of time varying sequences [29], which is an essential part for the remaining surgery duration estimation.

The method that was used in this thesis made use of the functions available in the *sklearn.neural\_network* package, with the following settings:

```
class sklearn.neural_network.MLPRegressor (hidden_layer_si-
zes=(100, ), activation='relu', *, solver='adam', alpha=0.0001,
batch_size='auto', learning_rate='constant', learning_rate_
init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_sta-
te=None, tol=0.0001, verbose=False, warm_start=False, mo-
mentum=0.9, nesterovs_momentum=True, early_stopping=False,
validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08,
n_iter_no_change=10, max_fun=15000
```

The full explanation of each attribute can be found in the Scikit-learn documentation [22]. The inputs used for this method were the frame properties as described in Table 2.3.

### 2.3.5. Decision Tree Regression

A decision tree is a more simplistic method which is based on conditional choices. A tree is constructed with nodes. At each node, the decision is made to either take the left or right node in the next layer, based on the condition of that node. This results in a final output (as can be seen in the example tree in Figure 2.4). A tree is constructed by taking a random predictor and condition and retraining it with the next frame from the training set, until the evaluation criteria converges. The advantage of decision trees is similar to that of the linear regression method; it creates explainable conditions for the remaining surgery duration. The tree can have the same result as a linear regression model if the predictors are linearly correlated to the remaining surgery duration. However, decision trees have the benefit that they do not need to confine the conditions on linearity, as conditions could reoccur in the tree causing the addition of non-linear correlations. The disadvantage of decision trees is that during the training, a random predictor is chosen as first condition. This could negatively influence the possibility of creating an optimal tree, because this sets the starting point for the tree [30]. The tree is trained, but will always retain this starting point. Another disadvantage is overfitting. Decision trees have the tendency to overfit, as the training algorithm has the ability to replicate the conditions in the tree [31]. This could show a better result than actually possible in practice. Lastly, decision tree regressors do not have the ability to use temporal information, so it is only able to use the information of the current frame.

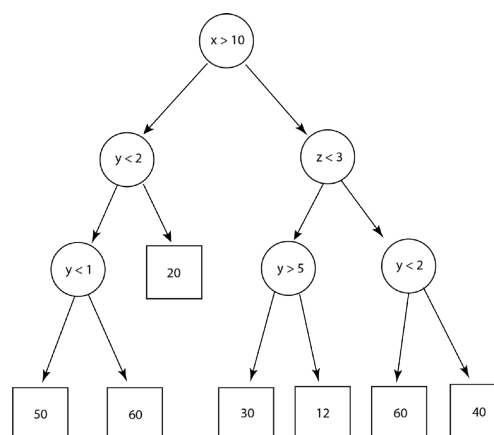


Figure 2.4: Graphical representation for an example of a decision tree

The decision tree regressor for this research was used from the *sklearn.tree* package with the following settings:

```
class sklearn.tree.DecisionTreeRegressor(*, criterion='mse',
splitter='best', max_depth=None, min_samples_split=2, min_sam-
ples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None,
random_state=None, max_leaf_nodes=None, min_impurity_de-
crease=0.0, min_impurity_split=None, presort='deprecated',
ccp_alpha=0.0
```

The full explanation of each attribute can be found in the Scikit-learn documentation [22]. The inputs used for this method were the frame properties as described in Table 2.3.

### 2.3.6. Random Forest Regression

The random forest regression model is a combination of multiple decision trees as explained in subsection 2.3.5. However, instead of using the result of one tree, the average of multiple trees is used. This solves the problem of decision trees where the starting point for constructing a decision tree could create a biased tree for specific types of datasets, as multiple trees are created with each a different starting value [30]. However, the other disadvantages of a decision tree still apply, such as the tendency to overfit and the inability of the use of temporal information.

The random forest regressor was retrieved from the *sklearn.ensemble* package with the following settings:

```
class sklearn.ensemble.RandomForestRegressor(n_estimators=100, *,
      criterion='mse', max_depth=None, min_samples_split=2, min_sam-
      ples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto',
      max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_
      split=None, bootstrap=True, oob_score=False, n_jobs=None, random_
      state=None, verbose=0, warm_start=False, ccp_alpha=0.0, max_sam-
      ples=None)
```

The full explanation of each attribute can be found in the Scikit-learn documentation [22]. The inputs used for this method were the frame properties as described in Table 2.3.

## 2.4. Novel method: Phase-Inferred with Nearest Neighbours and Dynamic Time Warping

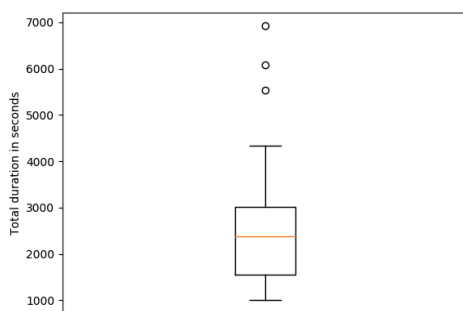


Figure 2.5: Boxplot with the total duration for each video in the Laparoscopic Cholecystectomy dataset in seconds

The phase-inferred method described in subsection 2.3.2 uses all the training videos to estimate a median or mean. However, this is not desirable as not all the videos are the same. As can be seen from Figure 2.5, the total duration of the videos differ. Some videos have a duration of 1000 seconds ( $\approx 16$  min), while others videos have a duration of over 6000 seconds ( $\approx 100$  min). If the average is used of all the videos, the total duration of the predictions will be around 3000 seconds ( $\approx 50$  min). Instead of using all the videos as  $N$  in Equation 2.2, a more effective approach would be by comparing the videos and only using the mean or median of similar videos (also called ‘nearest neighbours’). However, to compare datasets correctly, a

similarity metric is needed to be able to determine similar videos. Commonly used measures are Euclidian, Manhattan or Cosine distances [32–34]. But these distances only compare one frame to another frame at the same time index. The issue for time serie problems is that some videos can still be similar, but have a small shift in the starting points (or other key points) [35]. For that reason a different similarity measurement approach was used: Dynamic Time Warping [36]. Dynamic Time Warping is a dynamic approach used to compare time series and to find patterns between these series [36]. The main concept with Dynamic Time Warping is that all frames are used to measure the distance, instead of only the frame at the same time index.

Using a dynamic algorithm, the shortest path can be found. This is done by aligning similar-like points of the time series, so that the distance of each of these points can be calculated (as can be seen in Figure 2.6). The type of measurement between these distances can be any type of distance measurement, depending on the data and the usage. In this case, the Euclidian distance was used for all the frame properties as listed in section 2.1. The Dynamic Time Warping algorithm used in this research was based on FastDTW [37]. The  $k$  videos with the smallest distances were used as reference points for the mean or median durations for the phases instead of all videos. The choice was made of four videos, which was based on the results of multiple runs with different numbers ( $k=4, 5, \dots 10$ ) and the mean absolute error as evaluation metric (as explained in subsection 2.5.1). The pseudocode for retrieving the four similar videos using  $k$ -nearest neighbors is shown in algorithm 2. This method is called the novel method for the remainder of this thesis. While combining the nearest neighbour algorithm with Dynamic Time Warping is not new, it has however never been used for the estimation of the remaining surgery duration, to the authors knowledge. It is a novel method for this field.

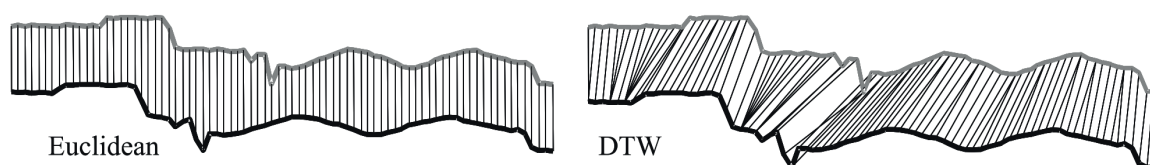


Figure 2.6: Euclidian distance alignment compared to Dynamic Time Warping [38]

---

**Algorithm 2:**  $k$ -nearest neighbours with Dynamic Time Warping

---

```

i = current frame number
distance_list =  $\emptyset$ 
for Video  $v_i$  in videolist_trainingset do
    distance = dtw(testvideo,  $v_j$ , i)
    distance_list.add((distance,  $v_j$ ))
end for
distance_list.sortOnDistance()
return First  $k$  videos from distance_list.values()

```

---

## 2.5. Evaluation

There are two parts for the evaluation; the evaluation of the technical aspects of the estimation model on a system level (subsection 2.5.1), and an evaluation of the added value of the intraoperative remaining surgery duration estimation system in the operating room workflow (subsection 2.5.2).

### 2.5.1. System level evaluation

The system level was evaluated using the mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE). Each method was evaluated using these measures and compared based on a leave-one-out cross validation [39], where the training set used the manual annotated phases for the estimation and the test video used the predicted phases. The evaluation was based on three parts; the complete dataset, grouped by surgery length, and split into quarters based on the duration. The grouping was done in short, medium and long surgeries, which contained respectively 25%, 50%, and 25% of the data. As seen in Figure 2.1, the short surgeries are videos which fall in the 25<sup>th</sup> percentile; the long surgeries are videos from the 75<sup>th</sup> percentile to the 100<sup>th</sup> percentile. The rest was grouped as medium surgeries. The split of the data for each video based on surgery duration was done into quarters. The quarters were based on dividing the surgery duration for each video in four and calculating the error of that part of the surgery. For example, for a video of 1000 seconds, the mean absolute error of the first quarter is the sum of the error of second 0 to second 250, divided by this duration. Using quarters to evaluate the estimation system has been used in previous research [17, 19], as it gives a clear distinction between the different parts of a surgery and simple to use for creating a comparison between surgeries.

#### Mean Absolute Error

The mean absolute error (MAE) describes the average error of the estimated remaining duration compared to the real duration of the surgery, with:

$$MAE = \frac{1}{n} \sum |t_r - t_p| \quad (2.4)$$

where  $t_p$  is the predicted remaining duration,  $t_r$  the real remaining duration, and  $n$  the number of surgeries in the dataset. It gives an indication of the error margin of the estimation system. The MAE is the most common used evaluation metric for intraoperative remaining surgery duration estimation systems [14, 16, 17], as it is a more natural and unambiguous error metric than other error metrics, like for example the root mean squared error [40]. However, the accuracy also depends on the total duration of the surgery, as a small MAE with a relative long surgery is significantly better than the same MAE with a short surgery. Also, the standard deviation of the MAE defines the reliability of system, as a small deviation defines an on average reliable prediction system. This, however, can only be used as a comparative metric for methods that use the same dataset, as it is independent of the surgery duration.

#### Mean Absolute Percentage Error

Another measurement method is the mean absolute percentage error (MAPE). As the MAE is independent of the surgery duration, the MAPE gives a mean error relative to the total surgery duration. The MAPE is defined as following:

$$MAPE = \frac{100}{n} \sum \left| \frac{t_r - t_p}{t_r} \right| \quad (2.5)$$



The MAPE shows the percentage of error relative to the real duration, giving an more straightforward metric to compare different datasets with different durations.

### Root Mean Squared Error

A common used evaluation metric which is comparable to the MAE is the mean squared error (MSE), with:

$$MSE = \frac{1}{n} \sum (t_r - t_p)^2 \quad (2.6)$$

The difference of the MSE compared to the MAE is that instead of taking the absolute value of the error, it takes the squared error. Outliers have a bigger influence on the MSE than in the MAE [40]. The MSE reports a better indication for the reliability of the system, while the MAE shows a more accurate overall performance as it does not take outliers into account as significantly as the MSE. As the MSE does not return a scaled error relative to the actual error, the root mean squared error (RMSE) was used, with:

$$RMSE = \sqrt{\frac{1}{n} \sum (t_r - t_p)^2} \quad (2.7)$$

## 2.5.2. Operating room workflow evaluation

The operating room workflow evaluation was based on two parts: an analysis of available data of the Spaarne Gasthuis hospital, and interviews done with current operating room staff, specifically the operating room program coordinators (ORPCs).

The available data contained all the surgeries from 2016 to 2019 of the Spaarne Gasthuis hospital. The Spaarne Gasthuis has three locations, Hoofddorp, Haarlem-North and Haarlem- South. They are abbreviated to HO, HAN, and HAZ respectively for the remainder of this report. The data contained timestamps for each surgery for multiple moments, such as the start and end of each surgery. This was used to analyse the following parts of the operating room workflow, specifically regarding the operating room schedule:

- **Turnover time:** The time from end of surgery for one patient and start surgery for the next patient.
- **Overtime:** The time an operating room is utilized after 16:30.
- **Undertime:** The time an operating room is not utilized before 16:30, after an existing surgery in the operating room.

The timestamps used for the analysis were the start and end surgical time. However, the turnover time included other durations such as intubation time. This increased the turnover time, however, this was due to the fact the start and end surgical times were the only available data for the analysis. Furthermore, only elective surgeries were included in the analysis for the undertime and the overtime. Non-elective surgeries do not follow the standard scheduling process and were therefore excluded.

The interviews were performed with current users of the scheduling system, the operating room program coordinators (ORPCs). The ORPCs are responsible for the daily schedule and changes this schedule according to any events occurring during the day. If a remaining duration estimation system would be used in practice, the ORPCs would be the main users of the system. The interviews were conducted to retrieve information from the current creators of an operating room schedule, information that would give more in-depth details about the

usefulness of an intraoperative estimation system in practice. The focus of the interview was understanding the current workflow and how an intraoperative remaining surgery duration system would be beneficial in this workflow. The results were also used to be able to discuss the results of the operating room data (turnover time, overtime, undertime) as previously described. The results of this interview were subjective, as they were based on the experience and opinion of each individual. However, it gave an initial starting point to better understand the usefulness and the requirements for such a system in the operating room workflow.

Each interview was conducted using the questions as shown in Table 2.5. A total of three participants were included in the interview process. The participants first received a brief explanation about this research and the main goal of the automatic remaining surgery duration system. Afterward, the questions were asked and recorded.

Table 2.5: Interview questions for the operating room program coordinators which evaluated the operating room schedule and the addition of an automatic intraoperative remaining surgery duration estimation system

| <b>Interview question</b>   |   |
|---|---|
| <i>Current operating room workflow</i>                                      |   |
| Q1  | How do you make sure that the schedule is correct during the day? What do you change?                                     |
| Q2  | Which information do you need during the day to be able to change the schedule properly?                                  |
| Q3  | How do you retrieve information about the progress of current surgeries?  |
| Q4  | Based on your experience, is the remaining surgery duration estimation of the surgeon during the surgery correct?         |
| Q5  | Are there types of surgeries where the surgery duration are more difficult to predict than others?                        |
| Q6  | Do you change the schedule during a surgery? If so, in which part of the surgery? (Q1, Q2, Q3, or Q4)                     |
| Q7  | In your experience, how much is the turnover time?  |
| Q8  | Is this necessary/needed, or too much/too little?   |
| Q9  | Is there a high occurrence of overtime for the operating room staff? (Daily, once a week,.. etc)                          |
| Q10   | What is the reason for overtime?  |
| Q11   | Is there a high occurrence of undertime for the operating rooms? (Daily, once a week,.. etc)                              |
| Q12   | What is the reason for undertime?   |
| <i>Addition intraoperative remaining surgery duration estimation system</i> |   |
| Q13   | Would you want to use an automatic surgery duration estimation system, and why (not)?                                     |
| Q14   | How would such a system help your daily work?   |
| Q15   | If you would use an automatic remaining surgery duration estimation system, what should the accuracy be to be acceptable? |
| Q16   | When during the surgery do you need the estimation? (Q1, Q2, Q3, Q4)  |

# RESULTS



This chapter shows the results of the methods as explained in chapter 2. Section 3.1 describes the results of the methods on a systematical level, as described more in detail in subsection 2.5.1. It used the results of the different methods to show the similarities and differences to compare the performance. Section 3.2 shows the results of the operating room workflow analyses; the data analysis and the expert interview results.

## 3.1. System level

Table 3.1 and Table 3.2 show the mean absolute error, root mean squared error and the mean absolute percentage error for each method for the Laparoscopic Cholecystectomy and Total Laparoscopic Hysterectomy dataset respectively, split into quarters based on the surgery duration and for the full duration. The best score, which is the methods with the lowest error, for each quarter is indicated in bold. These tables give an indication of the accuracy of the methods, as the goal for each method was to have the smallest error as possible. The naive method has been used in literature as baseline, as this is the current method used in practice. All methods had a smaller error for all the quarters compared to the naive method. As observed in Table 3.1 the naive method had the highest mean absolute percentage error of  $107\% \pm 109\%$ , while the next highest error was the decision tree with  $57\% \pm 44\%$  in the first quarter. Considering the full duration of the surgeries, the difference was even bigger, with the decision tree having a mean absolute percentage error of  $42\% \pm 31\%$ . This relative large difference is also seen in the Total Laparoscopic Hysterectomy dataset in Table 3.2. This shows that all methods performed better than the current method (the naive method) based on the mean absolute error, root mean squared error and mean absolute percentage error.

Based on the full duration, for both datasets the phase-inferred method using the median resulted in the smallest error with, for example, a mean absolute error of  $10.5 \pm 10$  for the Laparoscopic Cholecystectomy and  $9.6 \pm 6.5$  for the Total Laparoscopic Hysterectomy. However, looking at each quarter, the novel method has a smaller mean absolute error in the second quarter for the Laparoscopic Cholecystectomy and in the third quarter for the Total Laparoscopic Hysterectomy. It was a small difference, with both less than one minute. The similar results is understandable, as both relatively used the same method for estimating the remaining surgery duration.

Table 3.1: The mean absolute error (MAE) and root mean squared error (RMSE) with standard deviation in minutes and mean absolute percentage error (MAPE) with standard deviation as percentage for each method split into quarters of the total duration of each video for the **Laparoscopic Cholecystectomy** dataset

| Method                | First Quarter      | Second Quarter    | Third Quarter    | Fourth Quarter   | Full Duration      |
|-----------------------|--------------------|-------------------|------------------|------------------|--------------------|
| MAE (min)             |                    |                   |                  |                  |                    |
| Naive method          | 31.8 ± 18.8        | 31.8 ± 18.8       | 31.8 ± 18.8      | 31.8 ± 18.8      | 31.8 ± 18.8        |
| Phase-Inferred Mean   | 17.8 ± 15.6        | 15.1 ± 12.8       | 8 ± 7            | 5 ± 2.9          | 11.5 ± 9.6         |
| Phase-Inferred Median | <b>16.7 ± 17.1</b> | 14.1 ± 13.5       | <b>7.4 ± 7.1</b> | <b>3.7 ± 2.2</b> | <b>10.5 ± 10</b>   |
| Linear Regression     | 17.3 ± 15.8        | 14.4 ± 8.1        | 12.8 ± 6.6       | 10.5 ± 4.2       | 13.8 ± 8.7         |
| Multilayer Perceptron | 18 ± 15.7          | 15 ± 8.1          | 12.8 ± 7         | 8.2 ± 5.4        | 13.5 ± 9           |
| Decision Tree         | 22.1 ± 15.5        | 19.7 ± 11.8       | 15.6 ± 11.6      | 9.4 ± 8          | 16.7 ± 11.7        |
| Random Forest         | 21.2 ± 15.4        | 18.3 ± 10.3       | 15.3 ± 11.1      | 9.4 ± 8.4        | 16 ± 11.3          |
| Novel method Mean     | 19.1 ± 14.5        | 16.1 ± 9.5        | 14.2 ± 10.9      | 5.9 ± 5.4        | 13.9 ± 10.1        |
| Novel method Median   | 19.5 ± 15.1        | <b>13.2 ± 8.8</b> | 11.1 ± 7.6       | 4.6 ± 4          | 12.1 ± 8.9         |
| RMSE (min)            |                    |                   |                  |                  |                    |
| Naive method          | 31.8 ± 18.8        | 31.8 ± 18.8       | 31.8 ± 18.8      | 31.8 ± 18.8      | 31.8 ± 18.8        |
| Phase-Inferred Mean   | 18 ± 15.6          | 15.5 ± 13         | 8.6 ± 7.2        | 5.4 ± 3          | 11.9 ± 9.7         |
| Phase-Inferred Median | <b>16.8 ± 17.1</b> | 14.5 ± 13.7       | <b>7.9 ± 7.4</b> | <b>4.1 ± 2.4</b> | <b>10.8 ± 10.1</b> |
| Linear Regression     | 18.7 ± 15.9        | 15.4 ± 8          | 14.2 ± 6.4       | 12.3 ± 5.1       | 15.2 ± 8.8         |
| Multilayer Perceptron | 19.2 ± 15.7        | 16.2 ± 8          | 14.4 ± 6.9       | 10.5 ± 7         | 15.1 ± 9.4         |
| Decision Tree         | 24.8 ± 15.9        | 22.2 ± 13.5       | 19.8 ± 13.1      | 15 ± 11.1        | 20.5 ± 13.4        |
| Random Forest         | 23.1 ± 15.5        | 19.9 ± 11.3       | 18.9 ± 11.5      | 14.8 ± 11.4      | 19.2 ± 12.4        |
| Novel method Mean     | 20.3 ± 14.3        | 17.2 ± 9.7        | 15.7 ± 11.3      | 6.9 ± 6.5        | 15 ± 10.5          |
| Novel method Median   | 20.7 ± 14.8        | <b>14.2 ± 9</b>   | 12.3 ± 8         | 5.5 ± 4.8        | 13.2 ± 9.2         |
| MAPE (%)              |                    |                   |                  |                  |                    |
| Naive method          | 107% ± 109%        | 107% ± 109%       | 107% ± 109%      | 107% ± 109%      | 107% ± 109%        |
| Phase-Inferred Mean   | 54% ± 54%          | 45% ± 46%         | 23% ± 25%        | 15% ± 13%        | 34% ± 35%          |
| Phase-Inferred Median | <b>44% ± 41%</b>   | <b>36% ± 33%</b>  | <b>18% ± 17%</b> | <b>11% ± 9%</b>  | <b>27% ± 25%</b>   |
| Linear Regression     | 47% ± 42%          | 44% ± 38%         | 37% ± 27%        | 30% ± 19%        | 40% ± 31%          |
| Multilayer Perceptron | 50% ± 45%          | 46% ± 38%         | 35% ± 25%        | 23% ± 17%        | 38% ± 31%          |
| Decision Tree         | 57% ± 44%          | 51% ± 35%         | 37% ± 26%        | 24% ± 21%        | 42% ± 31%          |
| Random Forest         | 55% ± 43%          | 48% ± 31%         | 36% ± 25%        | 24% ± 22%        | 41% ± 30%          |
| Novel method Mean     | 54% ± 47%          | 48% ± 40%         | 37% ± 29%        | 16% ± 14%        | 39% ± 33%          |
| Novel method Median   | 54% ± 46%          | 37% ± 31%         | 27% ± 18%        | 12% ± 10%        | 33% ± 26%          |

Table 3.2: The mean absolute error (MAE) and root mean squared error (RMSE) with standard deviation in minutes and mean absolute percentage error (MAPE) with standard deviation as percentage for each method split into quarters of the total duration of each video for the **Total Laparoscopic Hysterectomy** dataset

| Method                | First Quarter    | Second Quarter    | Third Quarter    | Fourth Quarter   | Full Duration    |
|-----------------------|------------------|-------------------|------------------|------------------|------------------|
| MAE (min)             |                  |                   |                  |                  |                  |
| Naive method          | 29.1 ± 20.7      | 29.2 ± 20.8       | 29.2 ± 20.8      | 29.2 ± 20.8      | 29.2 ± 20.8      |
| Phase-Inferred Mean   | 15.2 ± 9         | 12.3 ± 7.8        | 8.5 ± 5.3        | 6 ± 3.4          | 10.5 ± 6.4       |
| Phase-Inferred Median | <b>15 ± 11</b>   | <b>12 ± 7.9</b>   | 7 ± 5.1          | <b>4.5 ± 2.2</b> | <b>9.6 ± 6.5</b> |
| Linear Regression     | 15.9 ± 9         | 13.2 ± 5.9        | 12 ± 4.5         | 9.9 ± 4.2        | 12.7 ± 5.9       |
| Multilayer Perceptron | 16.3 ± 8.4       | 12.4 ± 6.2        | 9.7 ± 4.1        | 9.8 ± 4.3        | 12 ± 5.8         |
| Decision Tree         | 19.8 ± 8.2       | 16.2 ± 5.3        | 10.4 ± 4         | 11.3 ± 5.6       | 14.4 ± 5.8       |
| Random Forest         | 18.4 ± 8.7       | 15.3 ± 5.6        | 9.9 ± 3.8        | 11.1 ± 5.6       | 13.7 ± 5.9       |
| Novel method - Mean   | 18.1 ± 10.5      | 13.9 ± 7.9        | 7.8 ± 3.8        | 6.5 ± 4          | 11.6 ± 6.6       |
| Novel method - Median | 16.5 ± 9.1       | 12.7 ± 7.2        | <b>6.4 ± 4.3</b> | 5.7 ± 3.8        | 10.3 ± 6.1       |
| RMSE (min)            |                  |                   |                  |                  |                  |
| Naive method          | 29.2 ± 20.7      | 29.2 ± 20.8       | 29.2 ± 20.8      | 29.2 ± 20.8      | 29.2 ± 20.8      |
| Phase-Inferred Mean   | 15.3 ± 8.9       | 12.6 ± 7.8        | 8.9 ± 5.2        | 6.5 ± 3.4        | 10.8 ± 6.3       |
| Phase-Inferred Median | <b>15.1 ± 11</b> | <b>12.3 ± 7.8</b> | 7.3 ± 5.2        | <b>4.9 ± 2.3</b> | <b>9.9 ± 6.6</b> |
| Linear Regression     | 18.3 ± 8.8       | 14.6 ± 5.2        | 13.9 ± 4.4       | 14.8 ± 6.2       | 15.4 ± 6.2       |
| Multilayer Perceptron | 18.3 ± 8.2       | 13.8 ± 6          | 11.4 ± 4         | 14.5 ± 6.3       | 14.5 ± 6.1       |
| Decision Tree         | 23 ± 8.5         | 19.3 ± 5.7        | 13 ± 4.2         | 16.6 ± 7.2       | 18 ± 6.4         |
| Random Forest         | 21.2 ± 8.7       | 18.2 ± 5.8        | 12.3 ± 4.2       | 16.3 ± 7.1       | 17 ± 6.5         |
| Novel method - Mean   | 19.5 ± 10.3      | 15.1 ± 8.2        | 8.6 ± 3.9        | 6.9 ± 4.1        | 12.5 ± 6.6       |
| Novel method - Median | 17.6 ± 9.1       | 13.9 ± 7.4        | <b>7 ± 4.3</b>   | 6.1 ± 4          | 11.2 ± 6.2       |
| MAPE (%)              |                  |                   |                  |                  |                  |
| Naive method          | 61% ± 57%        | 61% ± 57%         | 61% ± 57%        | 61% ± 57%        | 61% ± 57%        |
| Phase-Inferred Mean   | 31% ± 23%        | 26% ± 22%         | 18% ± 14%        | 13% ± 9%         | 22% ± 17%        |
| Phase-Inferred Median | <b>25% ± 14%</b> | <b>21% ± 12%</b>  | 13% ± 8%         | <b>9% ± 6%</b>   | <b>17% ± 10%</b> |
| Linear Regression     | 29% ± 14%        | 27% ± 19%         | 24% ± 15%        | 19% ± 10%        | 25% ± 14%        |
| Multilayer Perceptron | 32% ± 22%        | 26% ± 19%         | 19% ± 10%        | 18% ± 9%         | 24% ± 15%        |
| Decision Tree         | 38% ± 20%        | 32% ± 18%         | 19% ± 8%         | 21% ± 11%        | 28% ± 15%        |
| Random Forest         | 35% ± 20%        | 31% ± 19%         | 18% ± 8%         | 20% ± 12%        | 26% ± 15%        |
| Novel method - Mean   | 39% ± 32%        | 30% ± 25%         | 15% ± 11%        | 12% ± 8%         | 24% ± 19%        |
| Novel method - Median | 33% ± 25%        | 26% ± 22%         | <b>12% ± 10%</b> | 11% ± 8%         | 21% ± 16%        |

Figure 3.1 and Figure 3.2 are boxplots for the methods for each quarter using the mean absolute error of each video as data point. These show the range of the estimation error, to be able to see the deviation between the results for all the videos used in the datasets. It shows the number of outliers, and the range of these outliers. As can also be seen from Table 3.1 and Table 3.2, these box plots show that all the methods had a smaller error than the naive method. Specifically the fourth quarter, which is at the end of the surgery, the difference was relatively large. These figures also indicate that the range of the phase-inferred methods and the novel methods were smaller than the other methods.

Table 3.3 and Table 3.4 show the over- and underestimation for each method based on the same quarters. This describes the fraction of a quarter that had an underestimation and overestimation. To understand the results of the methods, it is useful to know when there is an error, if it is an under or over estimation. For example, the naive method for the Laparoscopic Cholecystectomy set had 78% of the error as an overestimation. This shows that in practice, a surgery will probably finish earlier than estimated. The phase-inferred method using the median had, on the full duration, a relative evenly distributed over- and underestimation (45% to 55% respectively). It is unclear what an optimal division would be in practice, that depends on the operating room workflow and the users. However, these tables can be used to find the method that is more fitting for each situation based on the over- and underestimation.

Table 3.5 and Table 3.6 also use the Laparoscopic Cholecystectomy and Total Laparoscopic Hysterectomy datasets respectively, however the results are grouped by the surgery duration: short, medium, and long surgeries. The difference between these tables and Table 3.1 and Table 3.2 is that it shows the behaviours of the methods for the different duration lengths. For example, at the Total Laparoscopic Hysterectomy dataset, the phase-inferred method had a better result for the medium group of videos, while the novel approach resulted better with long surgeries. This is explainable, as the phase-inferred method used all the videos to estimate the remaining surgery duration, which resulted in an average estimated duration over all the data. The novel approach only used a smaller group of videos, but more similar to the test video. Finding similarity in average videos is harder than longer videos, which are more kind of "outliers". However, this is not the case at the Total Laparoscopic Hysterectomy dataset, where the phase-inferred method and novel method had a relative similar estimation error ( $10.1 \pm 6.1$  and  $10.2 \pm 3$  respectively).



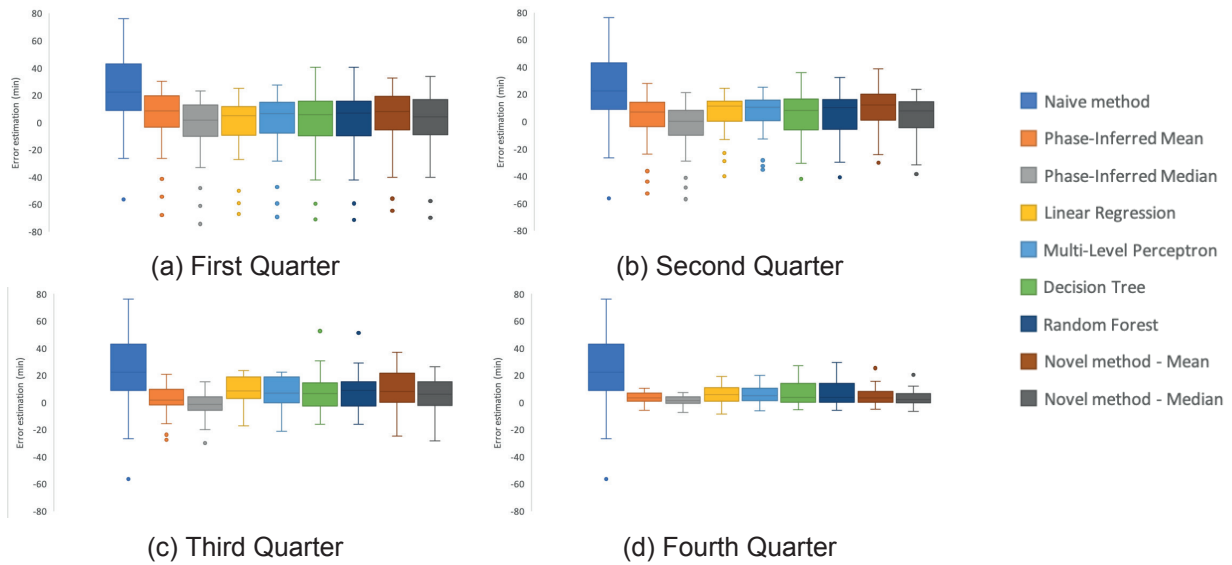


Figure 3.1: A boxplot representation of the mean absolute estimation error for the Laparoscopic Cholecystectomy dataset with for each quarter all the methods, with the estimation error in minutes.

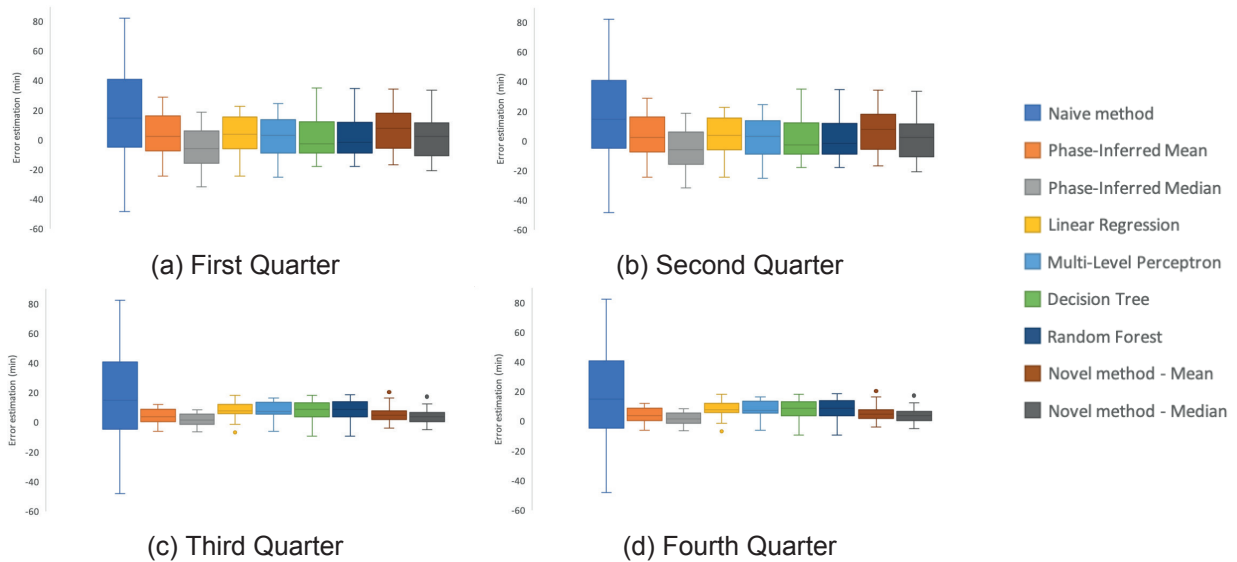


Figure 3.2: A boxplot representation of the mean absolute estimation error for the Total Laparoscopic Hysterectomy dataset with for each quarter all the methods, with the estimation error in minutes.

Table 3.3: The mean absolute (MAE) error for the underestimation and overestimation for each method grouped by surgery duration (see 2.1) for the **Laparoscopic Cholecystectomy** dataset. The fraction describes the division of under- and overestimation for each quarter.

|                       | First Quarter   |              | Second Quarter |              | Third Quarter |              | Fourth Quarter |              | Full duration |              |      |
|-----------------------|-----------------|--------------|----------------|--------------|---------------|--------------|----------------|--------------|---------------|--------------|------|
|                       | MAE (min)       | Fraction     | MAE (min)      | Fraction     | MAE (min)     | Fraction     | MAE (min)      | Fraction     | MAE (min)     | Fraction     |      |
| Naive method          | Underestimation | -22.2 ± 16.4 | 0.22           | -22.2 ± 16.4 | 0.22          | -22.2 ± 16.4 | 0.22           | -22.2 ± 16.4 | 0.22          | -22.2 ± 16.4 | 0.22 |
|                       | Overestimation  | 34.4 ± 18.8  | 0.78           | 34.4 ± 18.8  | 0.78          | 34.4 ± 18.8  | 0.78           | 34.4 ± 18.8  | 0.78          | 34.4 ± 18.8  | 0.78 |
| Phase-Inferred Mean   | Underestimation | -20 ± 23.1   | 0.34           | -15.9 ± 18.1 | 0.35          | -6.3 ± 8.3   | 0.40           | -2.7 ± 3.2   | 0.18          | -11.2 ± 13.2 | 0.32 |
|                       | Overestimation  | 15.5 ± 10    | 0.66           | 13.4 ± 8.4   | 0.65          | 6.6 ± 5.5    | 0.60           | 4.7 ± 3      | 0.82          | 10.1 ± 6.7   | 0.68 |
| Phase-Inferred Median | Underestimation | -22 ± 22.6   | 0.45           | -17 ± 17.3   | 0.48          | -7.7 ± 8.2   | 0.57           | -2.9 ± 3.1   | 0.29          | -12.4 ± 12.8 | 0.45 |
|                       | Overestimation  | 11.9 ± 8.2   | 0.55           | 9.9 ± 6.8    | 0.52          | 4.8 ± 4.1    | 0.43           | 3.1 ± 1.8    | 0.71          | 7.4 ± 5.2    | 0.55 |
| Linear Regression     | Underestimation | -20.2 ± 16.7 | 0.42           | -11.7 ± 10   | 0.23          | -6.3 ± 5.6   | 0.18           | -9.2 ± 2.4   | 0.30          | -11.8 ± 8.7  | 0.28 |
|                       | Overestimation  | 10.8 ± 8.5   | 0.58           | 11.8 ± 6.6   | 0.77          | 12.3 ± 6.9   | 0.82           | 11.7 ± 6.6   | 0.70          | 11.7 ± 7.2   | 0.72 |
| Multilayer Perceptron | Underestimation | -16.3 ± 16.6 | 0.39           | -11.4 ± 9.8  | 0.21          | -7.1 ± 6.8   | 0.31           | -4.3 ± 2     | 0.28          | -9.8 ± 8.8   | 0.30 |
|                       | Overestimation  | 12.3 ± 9.1   | 0.61           | 12.5 ± 6.8   | 0.79          | 13 ± 7.1     | 0.69           | 10.1 ± 7     | 0.72          | 12 ± 7.5     | 0.70 |
| Decision Tree         | Underestimation | -19.1 ± 16.1 | 0.43           | -13.9 ± 11.3 | 0.38          | -7.7 ± 6.9   | 0.43           | -2.7 ± 2.5   | 0.38          | -10.9 ± 9.2  | 0.41 |
|                       | Overestimation  | 18.4 ± 12.1  | 0.57           | 22 ± 15.9    | 0.62          | 22.6 ± 17.1  | 0.57           | 13.3 ± 12    | 0.61          | 19.1 ± 14.3  | 0.59 |
| Random Forest         | Underestimation | -19.5 ± 16.3 | 0.38           | -13.4 ± 11.7 | 0.33          | -7.9 ± 7.3   | 0.42           | -2.5 ± 2.5   | 0.34          | -10.8 ± 9.4  | 0.37 |
|                       | Overestimation  | 16 ± 9.9     | 0.62           | 17.9 ± 11.7  | 0.67          | 22.4 ± 15.9  | 0.58           | 12.7 ± 12.2  | 0.66          | 17.2 ± 12.4  | 0.63 |
| Novel method - Mean   | Underestimation | -16.3 ± 18.6 | 0.38           | -9.8 ± 9.1   | 0.22          | -4.7 ± 5.9   | 0.30           | -1.8 ± 2.1   | 0.18          | -8.1 ± 8.9   | 0.27 |
|                       | Overestimation  | 15.3 ± 8.4   | 0.62           | 15.1 ± 9.7   | 0.78          | 15.4 ± 11    | 0.70           | 5.9 ± 5.3    | 0.82          | 12.9 ± 8.6   | 0.73 |
| Novel method - Median | Underestimation | -16.9 ± 18.6 | 0.47           | -9.8 ± 11.3  | 0.42          | -5.5 ± 6.3   | 0.40           | -2.3 ± 2.3   | 0.27          | -8.6 ± 9.6   | 0.39 |
|                       | Overestimation  | 14.9 ± 8.8   | 0.53           | 11.8 ± 6.8   | 0.58          | 12 ± 7.6     | 0.60           | 4.4 ± 4      | 0.73          | 10.8 ± 6.8   | 0.61 |

Table 3.4: The mean absolute (MAE) error for the underestimation and overestimation for each method grouped by surgery duration (see 2.1) for the **Total Laparoscopic Hysterectomy** dataset. The fraction describes the division of under- and overestimation for each quarter.

|                       | First Quarter   |              | Second Quarter |              | Third Quarter |              | Fourth Quarter |              | Full duration |          |
|-----------------------|-----------------|--------------|----------------|--------------|---------------|--------------|----------------|--------------|---------------|----------|
|                       | MAE (min)       | Fraction     | MAE (min)      | Fraction     | MAE (min)     | Fraction     | MAE (min)      | Fraction     | MAE (min)     | Fraction |
| Naive method          | Underestimation | -20.4 ± 15.1 | 0.22           | -20.4 ± 15.1 | 0.22          | -20.4 ± 15.1 | 0.22           | -20.4 ± 15.1 | 0.22          | 0.22     |
|                       | Overestimation  | 33.2 ± 22    | 0.78           | 33.2 ± 22    | 0.78          | 33.2 ± 22    | 0.78           | 33.2 ± 22    | 0.78          | 0.78     |
| Phase-Inferred Mean   | Underestimation | -14.6 ± 10.2 | 0.34           | -9.7 ± 7.1   | 0.35          | -7.1 ± 6     | 0.40           | -3.4 ± 2.7   | 0.18          | 0.32     |
|                       | Overestimation  | 15.1 ± 8.6   | 0.66           | 10.7 ± 8.9   | 0.65          | 7.4 ± 5.3    | 0.60           | 5.9 ± 3.4    | 0.82          | 0.68     |
| Phase-Inferred Median | Underestimation | -17.7 ± 13   | 0.45           | -13.8 ± 8.8  | 0.48          | -7.7 ± 6.1   | 0.57           | -3.4 ± 2.4   | 0.29          | 0.45     |
|                       | Overestimation  | 9 ± 4.8      | 0.55           | 7.7 ± 5.1    | 0.52          | 4.8 ± 3.1    | 0.43           | 4.1 ± 2.3    | 0.71          | 0.55     |
| Linear Regression     | Underestimation | -22.6 ± 9.5  | 0.42           | -13.3 ± 5.6  | 0.23          | -6.8 ± 4.5   | 0.18           | -2.9 ± 1.6   | 0.30          | 0.28     |
|                       | Overestimation  | 9.1 ± 6.5    | 0.58           | 11 ± 6.9     | 0.77          | 12.1 ± 4.8   | 0.82           | 11.5 ± 5.9   | 0.70          | 0.72     |
| Multilayer Perceptron | Underestimation | -12.6 ± 10.2 | 0.39           | -9.8 ± 5.3   | 0.21          | -6.7 ± 4.6   | 0.31           | -2.6 ± 1.6   | 0.28          | 0.30     |
|                       | Overestimation  | 17.7 ± 10.6  | 0.61           | 10.6 ± 7     | 0.79          | 9.9 ± 3.7    | 0.69           | 11.5 ± 5.7   | 0.72          | 0.70     |
| Decision Tree         | Underestimation | -18.3 ± 9.2  | 0.43           | -13.4 ± 5.8  | 0.38          | -7.8 ± 4.5   | 0.43           | -3.9 ± 2.2   | 0.38          | 0.41     |
|                       | Overestimation  | 15.6 ± 8.9   | 0.57           | 16.3 ± 8.6   | 0.62          | 12.8 ± 5.6   | 0.57           | 16.1 ± 10.2  | 0.61          | 0.59     |
| Random Forest         | Underestimation | -17.6 ± 9    | 0.38           | -12.9 ± 6.3  | 0.33          | -7.5 ± 4.5   | 0.42           | -3.6 ± 2.2   | 0.34          | 0.37     |
|                       | Overestimation  | 13 ± 8.4     | 0.62           | 14.2 ± 8.6   | 0.67          | 11.8 ± 4.9   | 0.58           | 15.7 ± 10.4  | 0.66          | 0.63     |
| Novel method - Mean   | Underestimation | -10.3 ± 8.3  | 0.38           | -7 ± 4.8     | 0.22          | -5.4 ± 4.2   | 0.30           | -2.9 ± 2.5   | 0.18          | 0.27     |
|                       | Overestimation  | 17.4 ± 11.3  | 0.62           | 14.9 ± 8.4   | 0.78          | 7.9 ± 4      | 0.70           | 6.5 ± 4      | 0.82          | 0.73     |
| Novel method - Median | Underestimation | -12.3 ± 9.5  | 0.47           | -9.9 ± 5.9   | 0.42          | -5.1 ± 4.9   | 0.40           | -3.6 ± 2.2   | 0.27          | 0.39     |
|                       | Overestimation  | 13.6 ± 9.8   | 0.53           | 11.3 ± 8.4   | 0.58          | 5.3 ± 3.6    | 0.60           | 5.4 ± 4      | 0.73          | 0.61     |

Table 3.5: The mean absolute error (MAE) and root mean squared error (RMSE) with standard deviation in minutes and mean absolute percentage error (MAPE) with standard deviation as percentage for each method grouped by surgery duration (see 2.1) for the **Laparoscopic Cholecystectomy** dataset. Short: 25<sup>th</sup> percentile. Medium: 25<sup>th</sup> to 75<sup>th</sup> percentile. Long: 75<sup>th</sup> to 100<sup>th</sup> percentile

| Method                | Short             | Medium           | Long              | Complete           |
|-----------------------|-------------------|------------------|-------------------|--------------------|
| MAE (min)             |                   |                  |                   |                    |
| Naive method          | 46.9 ± 18.5       | 25.9 ± 15.6      | 25.7 ± 15.6       | 31.4 ± 18.8        |
| Phase-Inferred Mean   | 15.1 ± 5          | 5.8 ± 3.4        | 22.2 ± 11.4       | 11.5 ± 9.6         |
| Phase-Inferred Median | <b>11 ± 3.3</b>   | <b>4.7 ± 1.8</b> | 26.1 ± 11.5       | <b>10.5 ± 10</b>   |
| Linear Regression     | 15.1 ± 3.9        | 10.9 ± 3.8       | 19.9 ± 9          | 13.8 ± 8.7         |
| Multilayer Perceptron | 14.7 ± 2.7        | 10.2 ± 3.6       | 21 ± 10.7         | 13.5 ± 9           |
| Decision Tree         | 12.3 ± 5.1        | 15 ± 9.5         | 28 ± 5.3          | 16.7 ± 11.7        |
| Random Forrest        | 11.8 ± 5.2        | 14.3 ± 8.7       | 27.3 ± 5.7        | 16 ± 11.3          |
| Novel method Mean     | 13.1 ± 6.7        | 12.2 ± 6         | <b>19.6 ± 6.9</b> | 13.9 ± 10.1        |
| Novel method Median   | 11.5 ± 4.3        | 9.3 ± 3.4        | 20.7 ± 8.3        | 12.1 ± 8.9         |
| RMSE (min)            |                   |                  |                   |                    |
| Naive method          | 46.9 ± 19.2       | 25.9 ± 15.6      | 25.7 ± 15.6       | 31.8 ± 18.8        |
| Phase-Inferred Mean   | 15.7 ± 4.9        | 6 ± 3.4          | 22.7 ± 11.6       | 11.9 ± 9.7         |
| Phase-Inferred Median | <b>11.5 ± 3.3</b> | <b>4.9 ± 1.9</b> | 26.7 ± 11.7       | <b>10.8 ± 10.1</b> |
| Linear Regression     | 16.2 ± 4.1        | 12.3 ± 4.4       | 21.7 ± 9.2        | 15.2 ± 8.8         |
| Multilayer Perceptron | 16.3 ± 3.2        | 11.7 ± 4         | 23 ± 11           | 15.1 ± 9.4         |
| Decision Tree         | 16.3 ± 8.2        | 19.1 ± 11.7      | 30.6 ± 5.4        | 20.5 ± 13.4        |
| Random Forrest        | 15.8 ± 8.1        | 17.2 ± 9.8       | 29.8 ± 5.8        | 19.2 ± 12.4        |
| Novel method Mean     | 14.2 ± 6.7        | 13.5 ± 6.5       | <b>20.7 ± 6.8</b> | 15 ± 10.5          |
| Novel method Median   | 12.6 ± 4.1        | 10.5 ± 3.7       | 21.7 ± 8.3        | 13.2 ± 9.2         |
| MAPE (%)              |                   |                  |                   |                    |
| Naive method          | 238% ± 123%       | 65% ± 38%        | 29% ± 11%         | 107% ± 109%        |
| Phase-Inferred Mean   | 76% ± 31%         | 15% ± 11%        | 24% ± 7%          | 34% ± 35%          |
| Phase-Inferred Median | <b>56% ± 23%</b>  | <b>11% ± 4%</b>  | 29% ± 7%          | <b>27% ± 25%</b>   |
| Linear Regression     | 76% ± 26%         | 27% ± 10%        | <b>22% ± 5%</b>   | 40% ± 31%          |
| Multilayer Perceptron | 73% ± 21%         | 25% ± 10%        | 23% ± 7%          | 38% ± 31%          |
| Decision Tree         | 62% ± 30%         | 35% ± 20%        | 33% ± 4%          | 42% ± 31%          |
| Random Forrest        | 59% ± 30%         | 34% ± 19%        | 32% ± 4%          | 41% ± 30%          |
| Novel method Mean     | 65% ± 33%         | 30% ± 16%        | 23% ± 5%          | 39% ± 33%          |
| Novel method Median   | 57% ± 24%         | 23% ± 10%        | 24% ± 5%          | 33% ± 26%          |

Table 3.6: The mean absolute error (MAE) and root mean squared error (RMSE) with standard deviation in minutes and mean absolute percentage error (MAPE) with standard deviation as percentage for each method grouped by surgery duration (see 2.1) for the **Total Laparoscopic Hysterectomy** dataset. Short: 25<sup>th</sup> percentile. Medium: 25<sup>th</sup> to 75<sup>th</sup> percentile. Long: 75<sup>th</sup> to 100<sup>th</sup> percentile

| Method                | Short            | Medium           | Long              | Complete         |
|-----------------------|------------------|------------------|-------------------|------------------|
| MAE (min)             |                  |                  |                   |                  |
| Naive method          | 38.3 ± 28        | 26.4 ± 13.7      | 22.4 ± 15.5       | 29.2 ± 20.8      |
| Phase-Inferred Mean   | 14.9 ± 4         | 6.5 ± 3.7        | <b>10.1 ± 6.1</b> | 10.5 ± 6.4       |
| Phase-Inferred Median | <b>7.9 ± 3.3</b> | <b>5.7 ± 3</b>   | 15.8 ± 6          | <b>9.6 ± 6.5</b> |
| Linear Regression     | 14.1 ± 2.7       | 9.9 ± 2.4        | 14.4 ± 4.4        | 12.7 ± 5.9       |
| Multilayer Perceptron | 13.7 ± 3.3       | 9 ± 2.3          | 13.5 ± 3.2        | 12 ± 5.8         |
| Decision Tree         | 14.4 ± 4.5       | 12.5 ± 2.2       | 16.5 ± 2.6        | 14.4 ± 5.8       |
| Random Forest         | 13.9 ± 4.3       | 11.4 ± 2.1       | 15.9 ± 3          | 13.7 ± 5.9       |
| Novel method - Mean   | 15.7 ± 5.2       | 8.7 ± 2.6        | 10.2 ± 3          | 11.6 ± 6.6       |
| Novel method - Median | 12.8 ± 5.2       | 7.5 ± 1.8        | 10.8 ± 4.8        | 10.3 ± 6.1       |
| RMSE (min)            |                  |                  |                   |                  |
| Naive method          | 35.7 ± 30.1      | 27.9 ± 13.2      | 22.4 ± 15.5       | 29.2 ± 20.8      |
| Phase-Inferred Mean   | 15.3 ± 4.2       | 7 ± 3.7          | <b>10.6 ± 6.2</b> | 10.8 ± 6.3       |
| Phase-Inferred Median | <b>8.4 ± 3.3</b> | <b>5.8 ± 3.2</b> | 16.2 ± 6.2        | <b>9.9 ± 6.6</b> |
| Linear Regression     | 16.3 ± 2.4       | 12.7 ± 3.1       | 17.6 ± 3.7        | 15.4 ± 6.2       |
| Multilayer Perceptron | 16.2 ± 3.3       | 11.9 ± 2.8       | 16.1 ± 2.5        | 14.5 ± 6.1       |
| Decision Tree         | 18.7 ± 5.3       | 15.5 ± 3.1       | 20.2 ± 2.6        | 18 ± 6.4         |
| Random Forest         | 17.9 ± 5         | 14.3 ± 3.1       | 19.4 ± 2.8        | 17 ± 6.5         |
| Novel method - Mean   | 16.7 ± 5.4       | 9.8 ± 2.8        | 11.1 ± 3.2        | 12.5 ± 6.6       |
| Novel method - Median | 13.8 ± 5.5       | 8.3 ± 1.9        | 11.6 ± 4.9        | 11.2 ± 6.2       |
| MAPE (%)              |                  |                  |                   |                  |
| Naive method          | 95% ± 85%        | 53% ± 28%        | 30% ± 21%         | 61% ± 57%        |
| Phase-Inferred Mean   | 40% ± 15%        | 13% ± 9%         | <b>12% ± 6%</b>   | 22% ± 17%        |
| Phase-Inferred Median | <b>22% ± 11%</b> | <b>10% ± 4%</b>  | 19% ± 5%          | <b>17% ± 10%</b> |
| Linear Regression     | 39% ± 11%        | 18% ± 7%         | 18% ± 4%          | 25% ± 14%        |
| Multilayer Perceptron | 38% ± 13%        | 17% ± 6%         | 17% ± 3%          | 24% ± 15%        |
| Decision Tree         | 40% ± 13%        | 23% ± 5%         | 21% ± 3%          | 28% ± 15%        |
| Random Forest         | 38% ± 13%        | 21% ± 5%         | 20% ± 3%          | 26% ± 15%        |
| Novel method - Mean   | 43% ± 19%        | 17% ± 7%         | 13% ± 2%          | 24% ± 19%        |
| Novel method - Median | 36% ± 19%        | 14% ± 4%         | 13% ± 5%          | 21% ± 16%        |

Based on one of the video results for each dataset, a visualization is provided for each method in Figure 3.3 and Figure 3.4. The phase-inferred methods were relatively similar to the novel method, which is trivial as the core of both methods are the same. The difference can be seen when comparing Figure 3.3c and Figure 3.3e, where the phase-inferred method is more stable and constant, while the novel method has more deviations. This was due to the fact that the phase-inferred method is static, while the novel method is more dynamic. The phase-inferred method always used the same duration for the estimation of the next phases, as it used the mean or median of all the previous videos. The novel approach only used the similar videos, and continuously for each frame re-estimated the similarity between the videos. This gave the novel approach the ability to change and use other videos for the estimation during the progress of the video. Comparing these methods to the other methods, for example the decision tree in Figure 3.3h, the phase-inferred and novel methods had less deviation. This was also noticeable, for example, with the root mean squared error of the phase-inferred median method ( $10.8 \pm 10.1$ ), which was relatively similar to the mean absolute error ( $10.5 \pm 10$ ), while the decision tree root mean squared error ( $20.5 \pm 13.4$ ) was higher than the mean absolute error ( $16.7 \pm 11.7$ ). This was due to the fact that these methods estimation were only based on the current frame and did not use the results of the previous estimations, which was the case with the phase-inferred and novel methods. The only available data in a frame was the current phase of the frame, elapsed surgery duration on overall and in the current phase. This gave an indication on the progress, however it only used local information, and missed global data. On average, the error for these methods was relatively small, but the frequency of changes for an estimations was so large that it becomes questionable if the estimation could be used in practice.



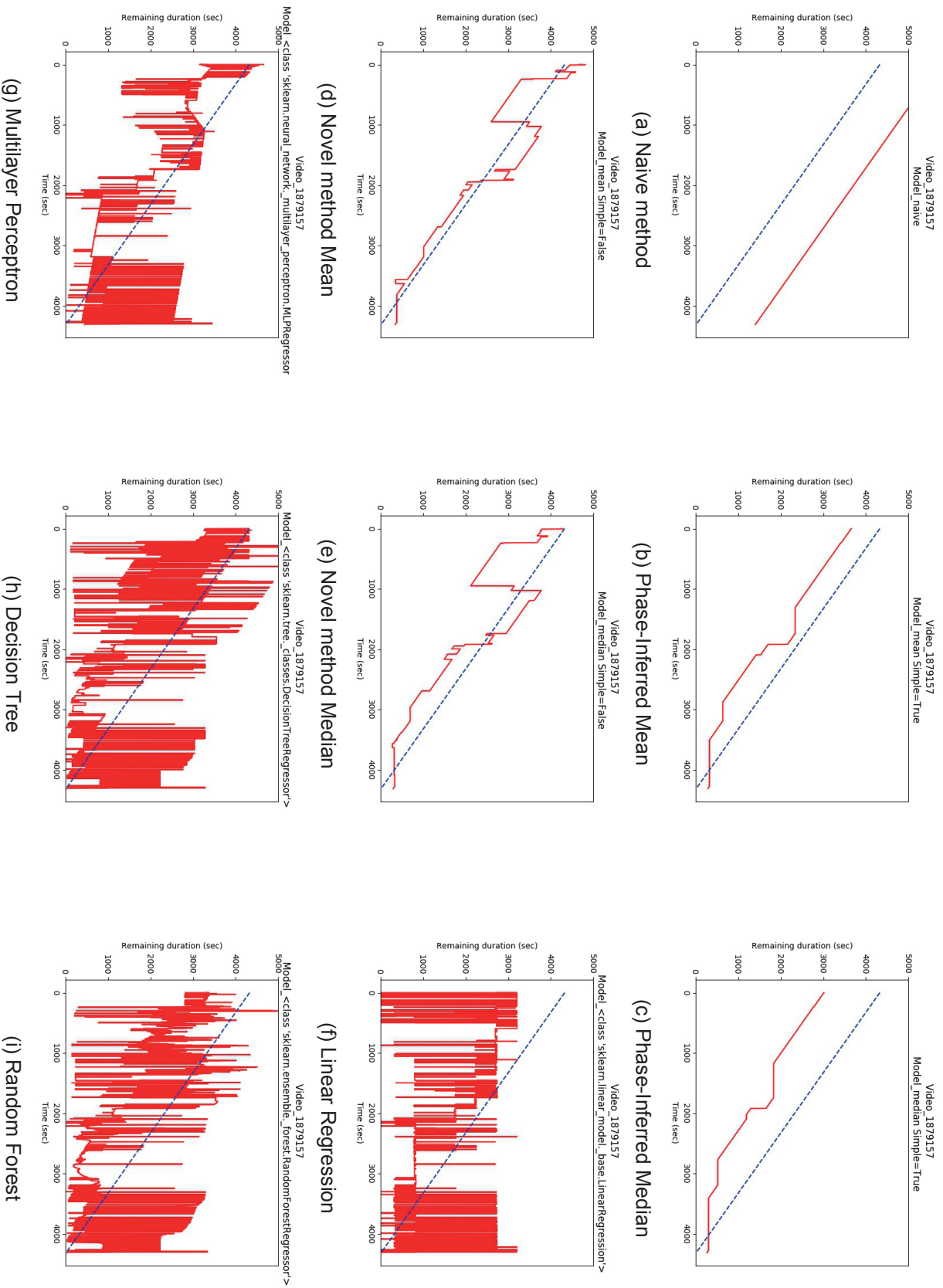
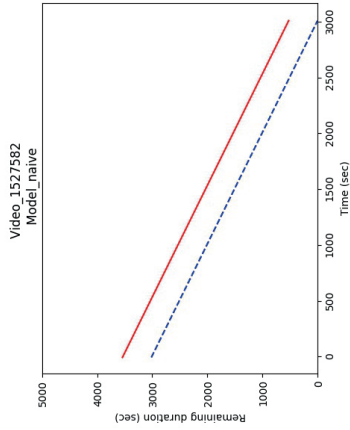
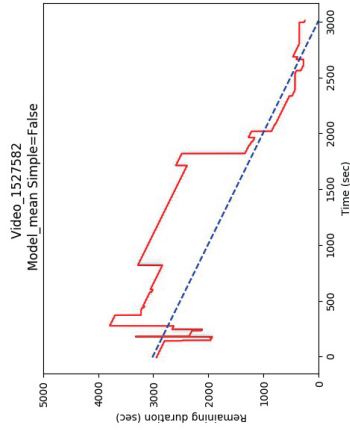


Figure 3.3: Remaining surgery duration estimation for video 1879157 from the Laparoscopic Cholecystectomy dataset for all the methods, with time and remaining duration in seconds. Blue dotted line: true remaining duration, Red unbroken line: estimated remaining duration.

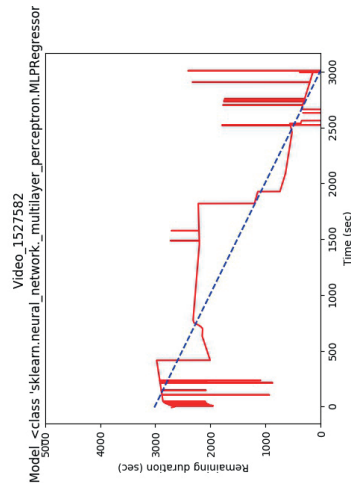




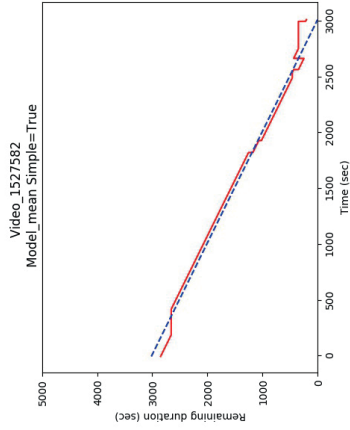
(a) Naive method



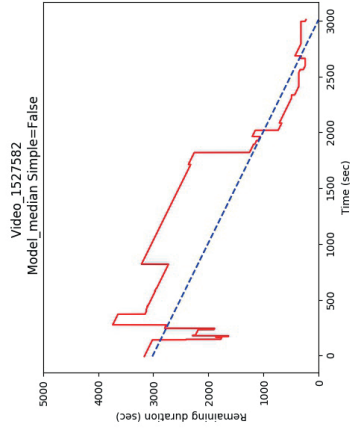
(d) Novel method Mean



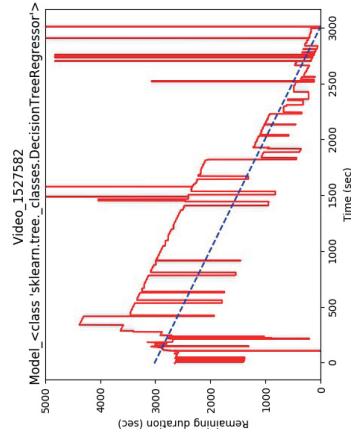
(g) Multilayer Perceptron



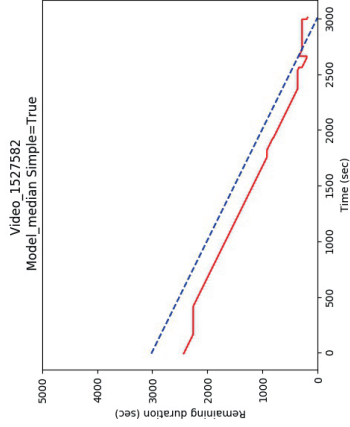
(b) Phase-Inferred Mean



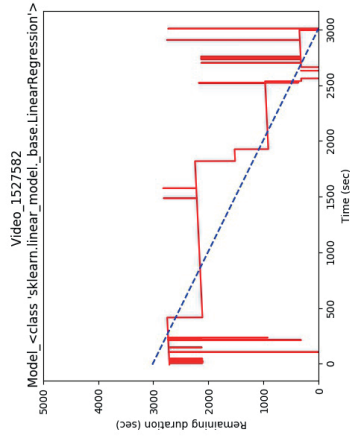
(e) Novel method Median



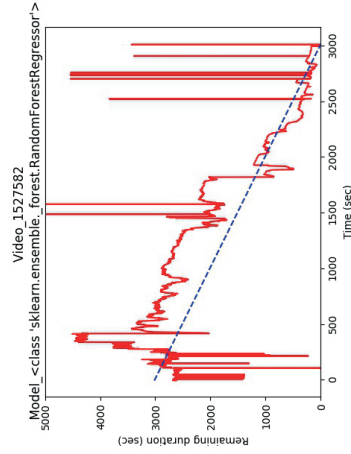
(h) Decision Tree



(c) Phase-Inferred Median



(f) Linear Regression



(i) Random Forest

Figure 3.4: Remaining surgery duration estimation for video 1527582 from the Total Laparoscopic Hysterectomy dataset for all the methods, with time and remaining duration in seconds. Blue dotted line: true remaining duration, Red unbroken line: estimated remaining duration.

## 3.2. Operating room workflow level

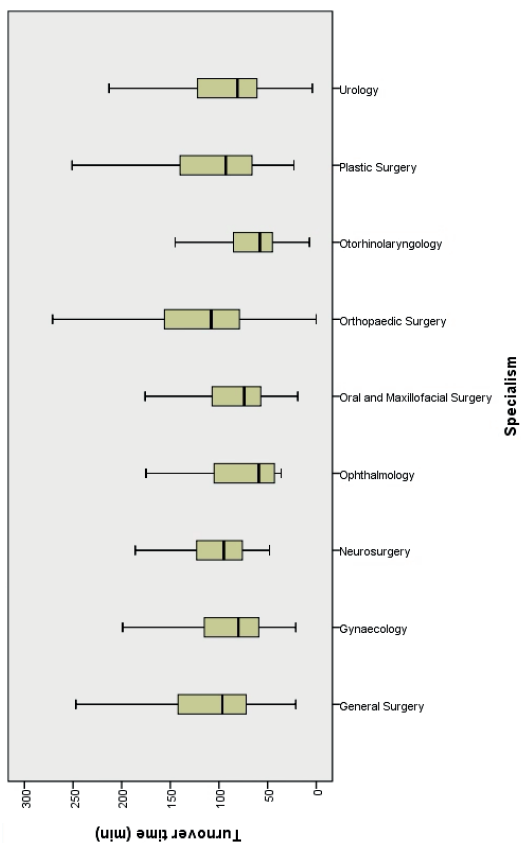
The operating room workflow was evaluated in two ways: a data analysis and expert interview. The data analysis was done on data of the Spaarne Gasthuis hospital in the Netherlands with the available surgical data from 2016 to 2019. The experts interviewed were operating room program coordinators (ORPSs).

### 3.2.1. Turnover time

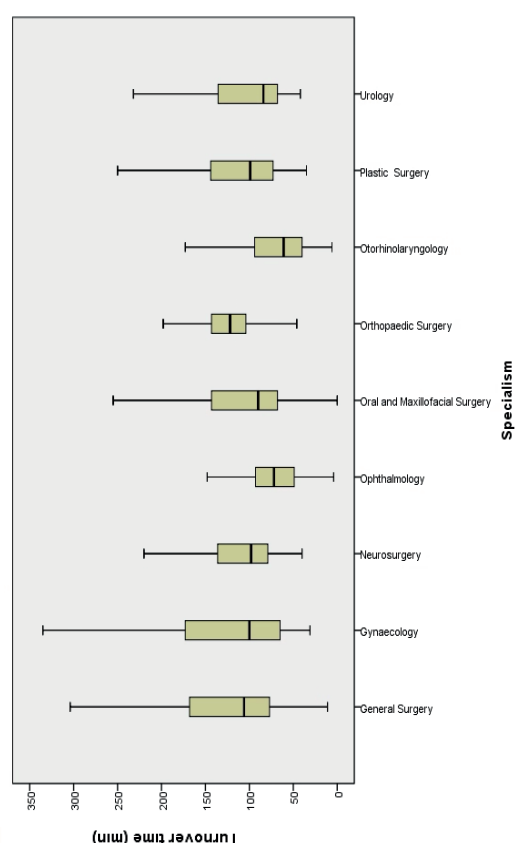
Table 3.7 and Figure 3.5 show the turnover time in minutes for the Spaarne Gasthuis hospital on overall and for each location. The turnover time is the time between two surgeries in the same operating room. However, as it was between end and start surgical time, the time included other parts such as intubation time. This means that the turnover time does not entirely indicate an unused operating room. The time after the last surgery was not included, as this would be an undertime or overtime. The turnover time is also grouped by specialism, which was done using the turnover time after each surgery. This, however, does not mean that the turnover time between two surgeries are both the same type of surgery, only the first surgery was used for the grouping. It shows the average turnover time after a specific specialism. This does not say that the type of surgery as said in the table directly was the cause of the turnover time, it could also be the next surgery. As can be seen, on overall Ophthalmology had on average the lowest turnover time (36 min) and General Surgery had the highest (119 min). Furthermore, location HAN had the shortest overall turnover time of 58 min minutes compared to the other locations, while HO and HAN had a similar number of surgeries. However, HAN had a smaller variety of types of surgeries, as they performed Ophthalmology surgeries more frequently ( $9262/19424 \approx 47\%$ ). Figure 3.5 is a boxplot representation of the turnover time for each specialism, showing not only the mean time, but also the range. Noticeable was that Otorhinolaryngology and Ophthalmology had on average a much smaller turnover time compared to the other surgeries.

Table 3.7: Descriptive statistics for the turnover time in minutes for each specialism from 2016 to 2019 at the Spaarne Gasthuis hospital locations and overall.

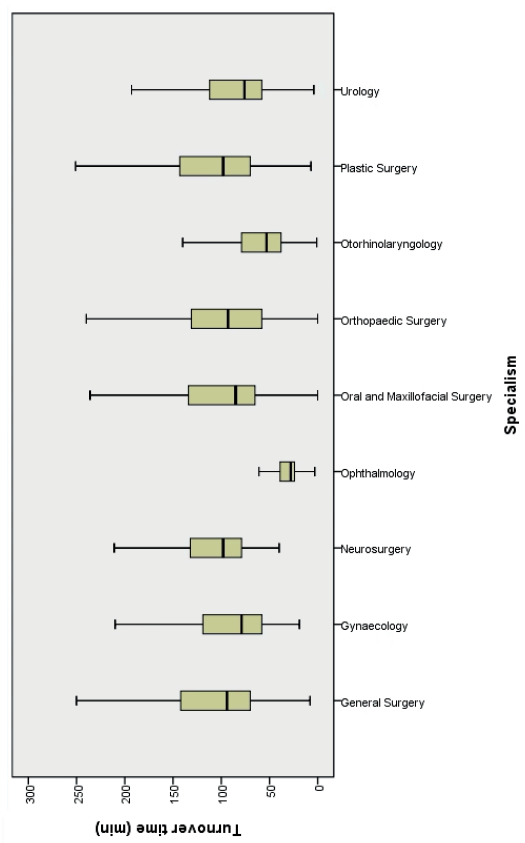
|                                | Overall |      | HO    |      | HAN   |      | HAZ  |      |
|--------------------------------|---------|------|-------|------|-------|------|------|------|
|                                | N       | Mean | N     | Mean | N     | Mean | N    | Mean |
| Overall                        | 48107   | 95   | 19148 | 113  | 19424 | 58   | 9535 | 131  |
| General Surgery                | 12875   | 119  | 5380  | 117  | 2387  | 84   | 5108 | 138  |
| Gynaecology                    | 3791    | 110  | 2313  | 96   | 612   | 73   | 866  | 172  |
| Oral and Maxillofacial Surgery | 1190    | 107  | 297   | 90   | 0     | -    | 893  | 113  |
| Neurosurgery                   | 1162    | 116  | 146   | 105  | 0     | -    | 1016 | 117  |
| Ophthalmology                  | 9355    | 36   | 8     | 78   | 9262  | 36   | 85   | 84   |
| Orthopaedic Surgery            | 9087    | 113  | 5749  | 134  | 3134  | 73   | 204  | 132  |
| Otorhinolaryngology            | 3793    | 65   | 1575  | 71   | 1726  | 56   | 492  | 74   |
| Plastic Surgery                | 3521    | 116  | 1331  | 113  | 1383  | 117  | 807  | 121  |
| Urology                        | 3333    | 96   | 2349  | 104  | 920   | 75   | 64   | 115  |



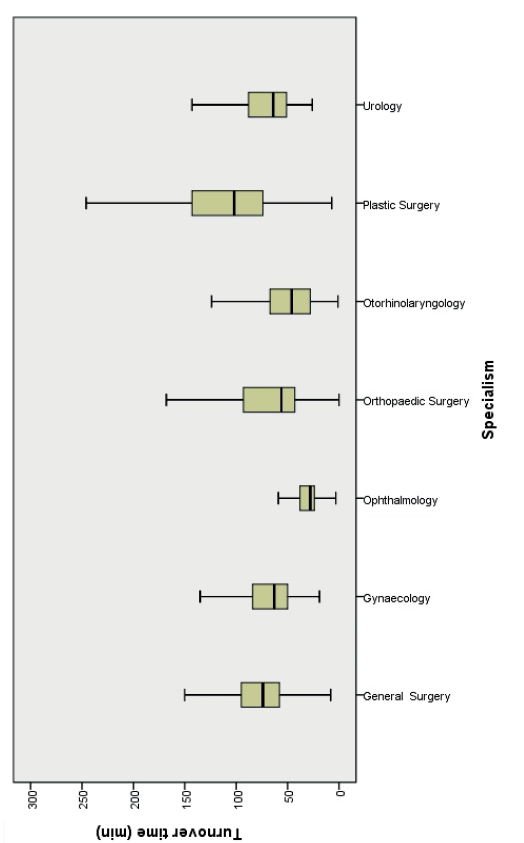
(a) Overall



(b) HO



(c) HAN



(d) HAZ

Figure 3.5: Boxplot representations for the turnover time grouped by specialism at the Spaarne Gasthuis hospital, on overall and for each location. Notice the different axis for each boxplot.

### 3.2. OPERATING ROOM WORKFLOW LEVEL

### 3.2.2. Overtime

The overtime was defined as the time after 16:30 used for surgery. The results only included elective surgeries, as non-elective surgeries do not follow the regular operating room schedule. Table 3.8 shows the overtime for the Spaarne Gasthuis hospital on overall and for each location grouped by specialism and Figure 3.6 is a boxplot representation of this overtime. From 2016 to 2019, there was a total overtime of 101572 minutes ( $\approx 1692$  hours). Overtime occurred for approximately 15% (1710/10818) of the surgeries that were the last surgery of the day in an operating room. On average, Plastic Surgery had the highest average overtime of 84 min. However, as General Surgery had significant more surgeries with overtime, it recounts for most of the overtime (852/1710  $\approx 50\%$ ). Another noticeable result was the frequency of surgeries that had overtime, where the HAN location was significant lower than the other locations. Also, the mean overtime for each specialism was significantly lower than the one of the HO and HAZ locations.

Table 3.8: Descriptive statistics for overtime of elective surgeries in minutes for each specialism from 2016 to 2019 at the Spaarne Gasthuis hospital locations and overall.

|                                | Overall |      | HO   |      | HAN |      | HAZ |      |
|--------------------------------|---------|------|------|------|-----|------|-----|------|
|                                | N       | Mean | N    | Mean | N   | Mean | N   | Mean |
| Overall                        | 1710    | 58   | 1014 | 57   | 107 | 20   | 589 | 65   |
| General Surgery                | 852     | 58   | 409  | 58   | 35  | 26   | 408 | 61   |
| Gynaecology                    | 131     | 79   | 84   | 81   | 1   | 8    | 46  | 79   |
| Oral and Maxillofacial Surgery | 26      | 67   | 9    | 57   | 0   | -    | 17  | 73   |
| Neurosurgery                   | 29      | 31   | 5    | 32   | 0   | -    | 24  | 31   |
| Ophthalmology                  | 20      | 29   | 0    | -    | 19  | 17   | 1   | 292  |
| Orthopaedic Surgery            | 325     | 48   | 299  | 50   | 16  | 10   | 10  | 72   |
| Otorhinolaryngology            | 61      | 35   | 33   | 34   | 10  | 22   | 18  | 45   |
| Plastic Surgery                | 123     | 84   | 41   | 96   | 19  | 26   | 63  | 94   |
| Urology                        | 143     | 50   | 134  | 53   | 7   | 14   | 2   | 5    |

### 3.2.3. Undertime

Undertime was defined as the time after the last surgery in an operating room, but before 16:30. This only included operating rooms that had at least one surgery that day. Table 3.9 shows the undertime for each location of the Spaarne Gasthuis hospital grouped by each specialism and Figure 3.7 is a boxplot representation. Undertime was empty operating room time that could be used for surgeries, however, it does not directly mean that this time would have been useful for a surgery. Other factors could be in play that are not visible in the results, as these results only show the time after the end of the last surgery. The highest undertime was found after Gynaecology surgeries, with an overall mean of 191 min. The lowest undertime for the specialisms was at General Surgery and Ophthalmology, both having on overall undertime of 85 min. The HAN location had on average the lowest undertime of 89 min, 30 min lower than the location with the highest undertime (HAZ).

Table 3.9: Descriptive statistics for the elective surgery undertime in minutes for each specialism from 2016 to 2019 at the Spaarne Gasthuis hospital locations and overall.

|                                | Overall |      | HO   |      | HAN  |      | HAZ  |      |
|--------------------------------|---------|------|------|------|------|------|------|------|
|                                | N       | Mean | N    | Mean | N    | Mean | N    | Mean |
| Overall                        | 9108    | 103  | 3756 | 100  | 3120 | 89   | 2231 | 129  |
| General Surgery                | 2654    | 85   | 1221 | 84   | 598  | 80   | 835  | 92   |
| Gynaecology                    | 968     | 191  | 474  | 125  | 127  | 127  | 367  | 300  |
| Oral and Maxillofacial Surgery | 366     | 117  | 63   | 133  | 0    | -    | 303  | 113  |
| Neurosurgery                   | 367     | 89   | 36   | 81   | 0    | -    | 331  | 90   |
| Ophthalmology                  | 869     | 85   | 0    | -    | 844  | 83   | 25   | 165  |
| Orthopaedic Surgery            | 1598    | 89   | 872  | 87   | 653  | 93   | 73   | 94   |
| Otorhinolaryngology            | 612     | 109  | 256  | 123  | 263  | 94   | 93   | 111  |
| Plastic Surgery                | 987     | 111  | 354  | 159  | 443  | 89   | 190  | 72   |
| Urology                        | 687     | 91   | 480  | 84   | 192  | 108  | 14   | 125  |

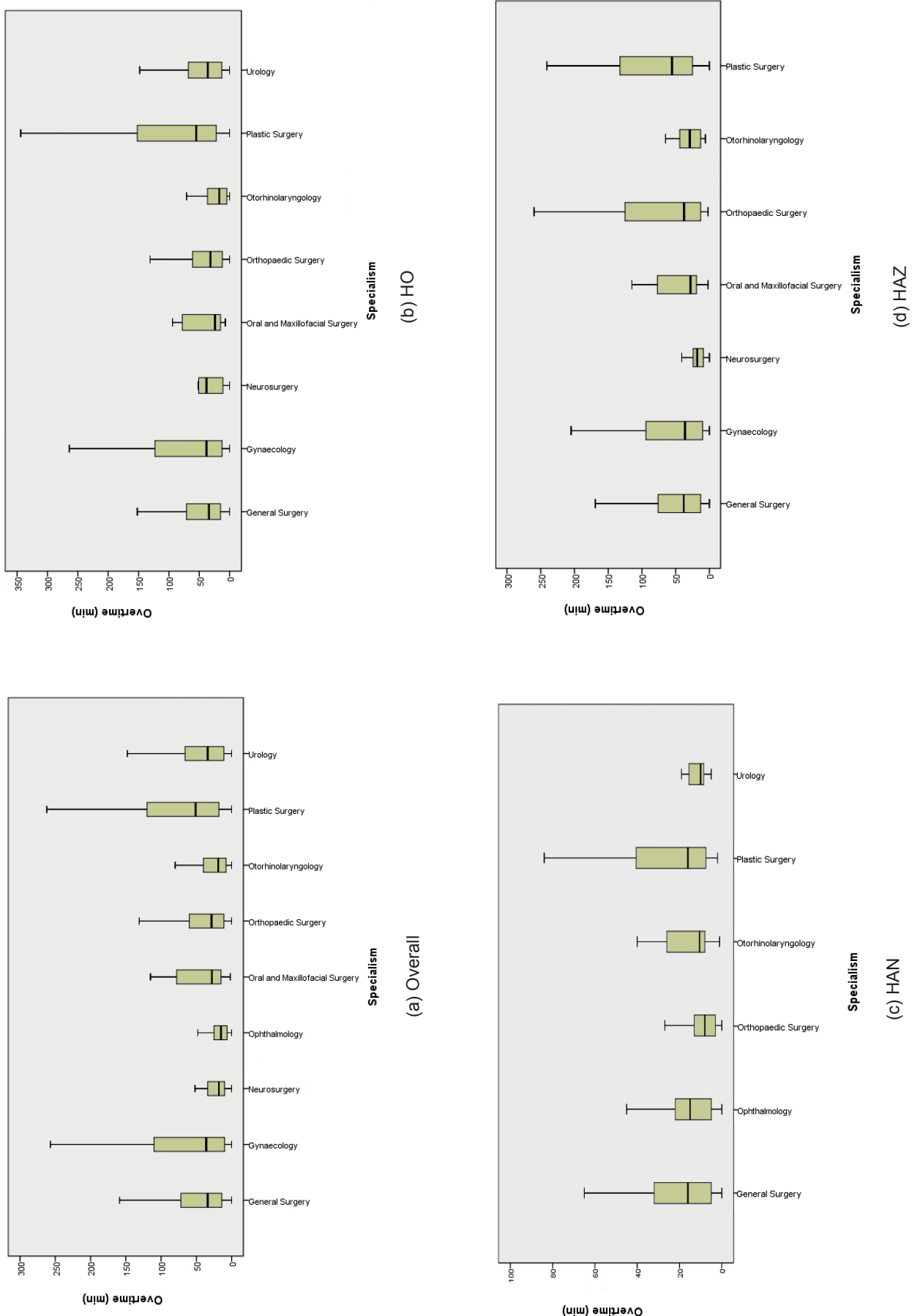


Figure 3.6: Boxplot representations for the overtime grouped by specialism at the Spaarne Gasthuis hospital, on overall and for each location. Notice the different axis for each boxplot.

### 3.2. OPERATING ROOM WORKFLOW LEVEL

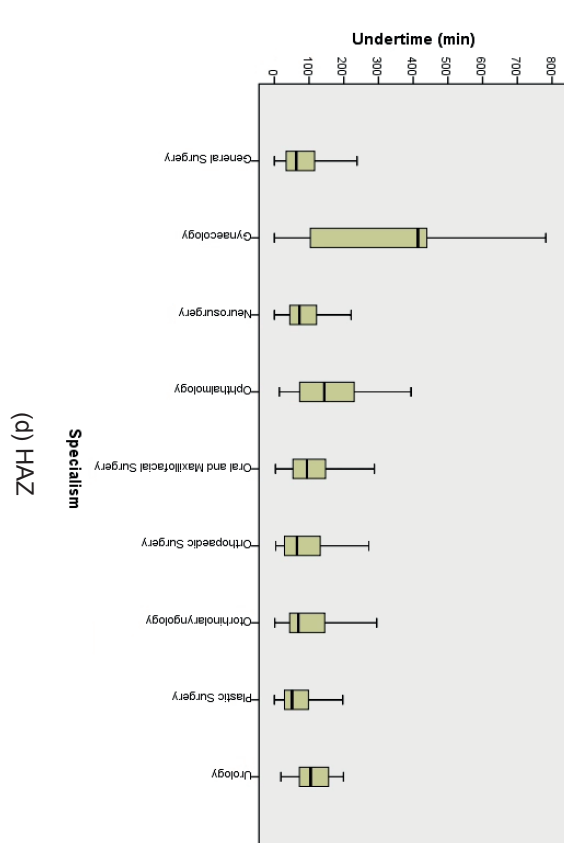
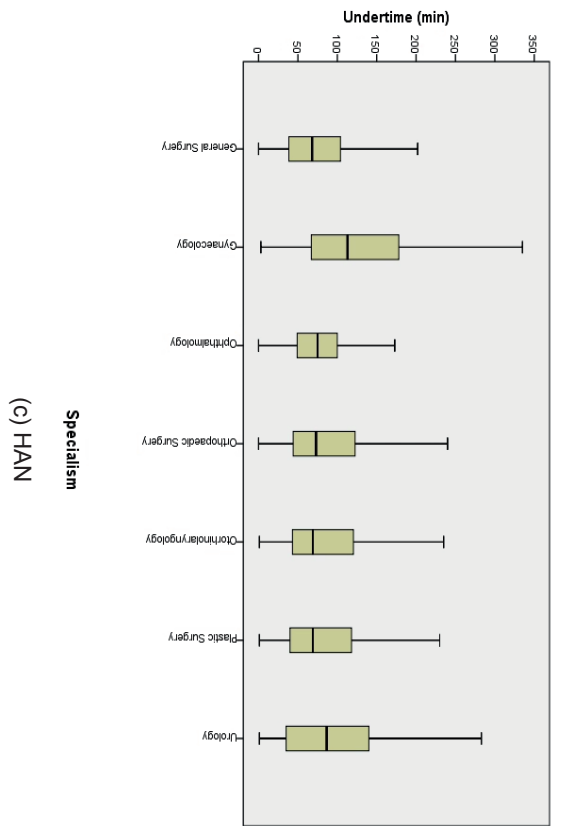
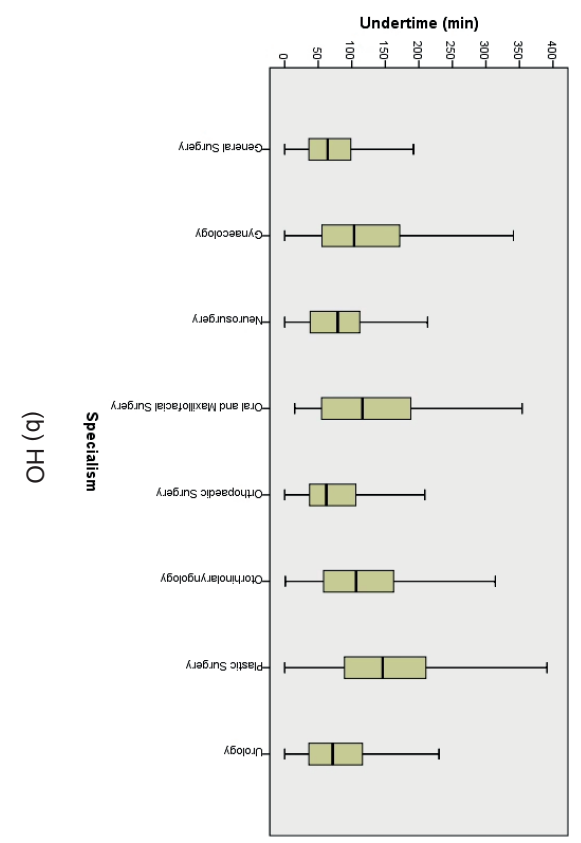
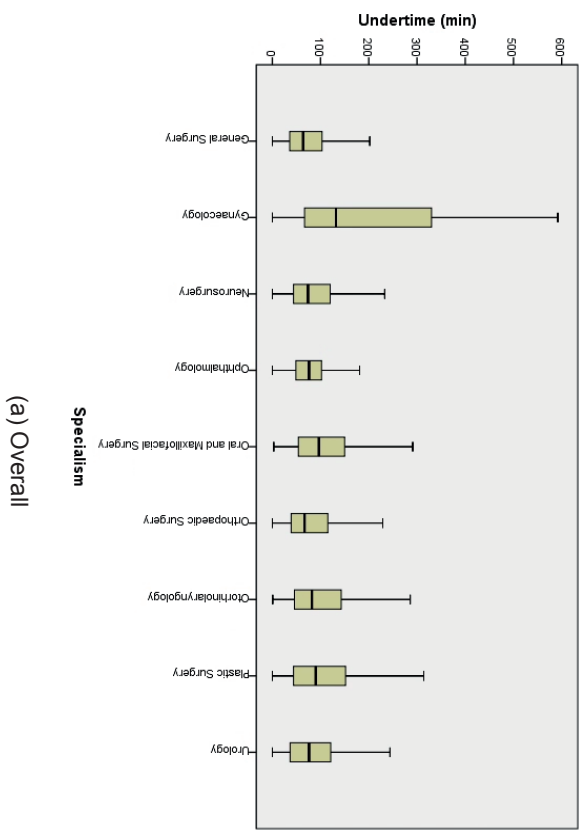


Figure 3.7: Boxplot representations for the undertime grouped by specialism at the Spaarne Gasthuis hospital, on overall and for each location. Notice the different axis for each boxplot.

### 3.2.4. Interview

The following results are a summary of each question for the interviews conducted as said in subsection 2.5.2. The results of each interview can be found in Appendix A. All the following results were based on their own opinion and experience.

**Q1. How do you make sure that the schedule is correct during the day? What do you change?**

The operating room schedules are created preoperatively using a system of the hospital. The ORPCs tune this schedule based on their own experience, notes of the ambulatory department, and information retrieved from and about the surgeon. Ambulatory is responsible for ordering the surgeries and indicates if the surgery would be different than an average surgery of that type. This gives the ORPCs information to be able to change the schedule accordingly. During the day, the ORPCs change the schedule based on the events happening in and at the operating room, such as delay or missing instruments.

**Q2. Which information do you need during the day to be able to change the schedule properly?**

Mainly the progress of the surgeries, and if there is a delay or not. One ORPC indicated that she needed to know the whereabouts of everyone of the medical staff, as this gives an indication of the progress. For example if an operating room assistant is going to get a coffee, she knew that the surgery is finished. All ORPCs need to be noticed by the operating room staff if there is a complication.

**Q3. How do you retrieve information about the progress of current surgeries?**

One ORPC tries to predict the durations of surgeries based on sight, while the other two use a more reactive method. The ORPC uses sight to understand the current progress, based on experience, to know if there could be a delay or not. All the ORPCs try to get an estimation of the surgeon and the operating room staff during the surgery but try to limit this to a maximum of once during a surgery, as this distracts the medical staff.

**Q4. Based on your experience, is the remaining surgery duration estimation of the surgeon during the surgery correct?**

All the ORPCs indicated to be hesitant about this estimation trustworthiness due to personal experience and knowledge. As the ORPCs know the surgeons and their ability to estimate the remaining duration (as some over or underestimate), they use the estimation of the surgeon more as an advice.

**Q5. Are there types of surgeries where the surgery duration is more difficult to predict than others?**

There are surgeries that are harder to estimate, mostly the longer surgeries than average. These surgeries can deviate a lot and have impact on the final schedule. The shorter ones are simpler to schedule. However, all ORPCs indicated that the remaining surgery estimation is hard to predict, as there are many factors influencing this prediction.

**Q6. Do you change the schedule during a surgery? If so, in which part of the surgery? (Q1, Q2, Q3, or Q4)**

There was no specific part of a surgery that the ORPCs use where they change the schedule, it is mostly based on the events happening in the operating room. The changes are made continuously, at any moment. For example, if the surgery started later than scheduled, this delay is automatically added in the schedule and changed accordingly.

**Q7. In your experience, how much is the turnover time?**

About 20 to 30 minutes, but they all indicated that the turnover time deviated significantly and frequently. It could be 15 min, but easily also a hour.

**Q8. Is this necessary/needed, or too much/too little?**

The turnover time deviates significantly, where multiple reasons could be the cause of the deviation. One ORPCs opinion was that the key factor for this is the anesthesiologist, as he/she needs to change between rooms. This causes delay for the change of patients. The other ORPCs had another opinion; the main reason for the deviation in turnover time are the patients. Especially calling the next patient and make sure the patient is ready for the surgery. The surgery duration was not the issue for the turnover time, it is the surgeons who most of the time need to wait on the patients to be ready. One of the main reasons of ending with a different schedule than preoperatively created is the turnover time.

**Q9. Is there a high occurrence of overtime for the operating room staff? (Daily, once a week,.. etc)**

Overtime occurred to their knowledge frequently, but not for a long duration. About once a week.

**Q10. What is the reason for overtime?**

The cause for overtime differed; it could be a delay in a surgery, but it could also be a logistic problem such as a missing instrument. It is an unavoidable issue that the ORPCs always anticipated and were prepared for it.

**Q11. Is there a high occurrence of undertime for the operating rooms? (Daily, once a week,.. etc)**

The ORPCs almost never observed undertime, no significant undertime where another surgery could had been performed.

**Q12. What is the reason for undertime?**

One ORPC explained that if there would be time for a surgery, a surgery would be scheduled. Another ORPC explained that undertime could occur when there is a cancellation and it would not be possible to call another patient.

**Q13. Would you want to use an automatic surgery duration estimation system, and why (not)?**

One ORPC had no interest in an estimation system, as it would not be possible to create such a system. There would be too many factors that needed to be taken into account, something a machine would not be able to do correctly. The other ORPCs would think that they would use such a system, but did not know why and how it would be useful directly.

**Q14. How would such a system help your daily work?**

It was not clear how such a system would help with the scheduling. Especially next to their current task, it seemed similar to check the current progress on a system compared to looking into the operating room. One of the concerns was the need to continuously check the estimation system to be able to use it correctly.



**Q15. If you would use an automatic remaining surgery duration estimation system, what should the accuracy be to be acceptable?**

The desired accuracy of the ORPCs was about 10 to 15 minutes.

**Q16. When during the surgery do you need the estimation? (Q1, Q2, Q3, Q4)**

Idealistic the system would estimate this at the third quarter of a surgery, as the beginning and end of surgery are not useful.

# DISCUSSION



This chapter discusses the used methods and results. It reflects on the feasibility of using an automatic intraoperative remaining surgery duration estimation system in the operating room workflow, based on the technical aspects of such a system and how such a system could be useful. The main goal of this research was to create an intraoperative remaining surgery estimation system to improve the operating room schedule. This was done by creating such a system and evaluating the system in two ways; quantitatively (system level and operating room workflow analysis) and qualitatively (expert opinion). Section 4.1 discusses the system results. Section 4.2 examines the current operating room workflow based on the analysis of previous operating room schedules and the results of the interviews. Furthermore, Section 4.3 discusses the system developed in this research based on all the results (system, operating room data analysis, and interviews) and how such a system could be implemented in the operating room workflow. Lastly, some future research suggestions are made to improve this research and to continue on in section 4.4.

## 4.1. System evaluation

Starting at the system level, the results show that multiple methods could be used to estimate the remaining surgery duration based on the surgical phases. The methods outperformed the naive approach based on the estimation error. The phase-inferred and novel method were the best performing methods compared to the other methods, and the performance difference between the phase-inferred and novel method was relatively small. Based on the large deviation observed at the linear regression, multilayer perceptron, decision tree, and random forest methods, it could be said that these methods are not usable in practice. As this research was focussed on creating a system that is feasible in the operating room workflow, a continuous changing estimation with a large difference would make the system unusable. The phase-inferred and novel methods are therefore the only methods that showed potential for the operating room workflow.

Comparing the phase-inferred and novel methods, the phase-inferred method showed more promising results on the average surgeries, while the novel method gained strength in performing better with outliers (e.g. longer surgeries than average). This is understandable, as the novel method tries to identify similar surgeries instead of using all the surgeries for the estimation. The ability of only using similar surgeries has the advantage to create a more dynamic estimation, as each new estimation would recalculate the similarity and in this way being able to change the estimation if a significant event occurs. The phase-inferred method has a fixed duration for each phase, which causes the method to only use the current phase to change the estimation, while the deviation of the current phase could also influence the duration of the other phases. An example could be that a surgeon is slower than average, which would increase the duration of all the phases, not only the current phase. The novel method is situation-specific, while the phase-inferred method is more fixed.

However, only using similar surgeries for the estimation comes with a cost, as it uses a smaller number of surgeries to estimate the surgery. This means that the similarity distance is a key

factor. This distance will eventually dictate which surgeries are used for the estimation, which is not always possible to be calculated correctly. For example, at the beginning of the surgery, many surgeries will seem similar. Based on this, the method will use a "random" group of surgeries for the estimation, which could cause a large error at the start of the surgery. As can be seen in Figure 3.4d, the fluctuation at the start of the surgery was large. It could be argued that this is an insignificant issue, as for scheduling purposes the remaining duration of the surgery at the beginning is less of an importance. However, it is something to be aware of, as this does influence the evaluation measurements as described in subsection 2.5.1.

Comparing the results of the two datasets, one notable point is the similar mean absolute error. The surgery duration of the Total Laparoscopic Hysterectomy was on average longer than the Laparoscopic Cholecystectomy. The Laparoscopic Cholecystectomy had an average duration of 70 min and the Total Laparoscopic Hysterectomy of 107 min. However, the mean absolute error was similar for both procedures on the full duration of the surgeries, about 10 min. As can be seen from the mean absolute percentage error, the methods performed better on the Total Laparoscopic Hysterectomy dataset compared to the Laparoscopic Cholecystectomy dataset. One of the explanations could be that the distribution of the surgery durations in the datasets was more evenly distributed in the Total Laparoscopic Hysterectomy than the Laparoscopic Cholecystectomy. This caused the outliers to be closer to the average than that of an unevenly distributed dataset.

One of the benefits of the system is scalability. The system was created on the Laparoscopic Cholecystectomy dataset and reused for the Total Laparoscopic Hysterectomy dataset. In general, no significant changes had to be made for the system to create an estimation for a new dataset, as the use of phases were the same. The Total Laparoscopic Hysterectomy dataset performed even better than the Laparoscopic Cholecystectomy dataset. However, the input has to be in the same structure, as it uses the phases to create an estimation. The phase-inferred and novel methods need to have sequential phases to be able to create an estimation, while the other methods use the phases categorically. One issue occurred at the Total Laparoscopic Hysterectomy dataset with the morcellation retrieval and transvaginal retrieval phases, where only one of the phases can occur, both at the same sequence location. That was solved by giving both phases the same "sequential" number. This caused the methods to have no distinction between the phases, creating a "retrieval" phase. This resulted in a loss of information for the system, however, based on the results it did not create a relatively bad performance. It is not recommended to merge phases, as this decreases the ability of the phase-inferred and novel methods to estimate the remaining surgery duration correctly earlier on the surgery, and the other methods will decrease in estimation performance in overall. It could be argued that the system should, in this case, create two types of surgeries: a Total Laparoscopic Hysterectomy with morcellation retrieval, and one with a transvaginal retrieval. However, this was not possible as the dataset was not large enough. Using a larger dataset, the novel method could dynamically use only videos that have the same phases, so only the one with for example the morcellation retrieval.

Furthermore, comparing the results to other systems and methods in literature based on the mean absolute percentage error, the system in this research performed similarly or better. Bodenstedt et al. [19] created a comparison with their methods and the methods of Twinanda et al. [17]. The best performing methods from Bodenstedt et al. and Twinanda et al. resulted on overall in a mean absolute percentage error of 23% and 25% respectively, and the phase-inferred method using the median resulted in a mean absolute percentage error of 27% for the Laparoscopic Cholecystectomy dataset and 17% for the Total Laparoscopic Hysterectomy dataset. Comparing the mean absolute percentage error during the progress of the surgeries, Bodenstedt et al. resulted in 17% for Q3 and 12% for Q4. This showed that the methods in this research could perform similar or better to other known methods in literature, on overall and

for the second part of the surgeries. However, it was not tested on the same dataset, making a direct comparison rather difficult.

Another noticeable point about the methods used in this research is the need for data in the datasets. The novel method tries to find similar videos to be able to estimate a remaining surgery duration. The datasets used in this research consisted of about 35 videos each. Increasing the number of videos could increase the number of similar videos, making the method possibly more effective. As the novel method continuously re-evaluates the similarity between all the videos and only uses a  $k$  number of videos, the estimation will be more accurate if there are more similar videos in the dataset. This is not the case with the phase-inferred method, as this method uses all the videos to estimate a mean duration for each phase. Adding more videos will change the mean, however, that will converge to a certain point. This will result in a relative same estimation for all the surgeries, as the estimation for the next phases is fixed. The novel approach has the benefit to be able to estimate the remaining surgery duration for outliers and to be able to increase the accuracy of the system after every new surgery added in the dataset. The problem, however, with increasing the data, is the computation time needed for the method. As the novel method needs to calculate the similarity at each frame for each video, the computation time increases significantly when adding more data. This could be solved by changing the current DTW to a more scalable method.

## 4.2. The current OR workflow

The current OR workflow was analyzed based on operating room data from 2016 to 2019, and interviews with experts. The analysis gave an indication of the current durations of the operating room workflow. It showed that there is room for improvement to decrease these durations, such as the turnover time. The interviews gave information about the scheduling process and what causes the non-surgical durations, such as overtime. It also assessed the acceptance of the ORPCs for using an estimation system, the need for such a system from a user perspective.

To start, the turnover time showed an average overall duration of 95 minutes, which differed for each specialism. One important notice in the results is that the turnover time in this research was defined as the time between the end of a procedure and the start of the next one. The turnover time therefore also included intubation time, cleaning the operating room, and other factors that made it not possible to use the operating room for a new surgery. For example, the average intubation time at the beginning of surgery was on average 9 minutes and at the end of a surgery 6 minutes, which is part of the turnover time as shown in the results. However, using start and end intubation time to analyze the turnover time was not possible, as not all the surgeries had an intubation time, and relatively much data was missing. Furthermore, other time stamps were missing such as cleaning the operating room, as these timestamps were not available in the dataset. This was also part of the turnover time as presented in the results. However, these results indicated the difference between the specialisms, the locations, and an estimate of the overall turnover time.

Furthermore, the turnover time between the locations showed a significant difference. Location HAN had on overall a turnover time of 58 min, while the others had a turnover time of more than 100 min. An explanation for this is that HAN performed many Ophthalmology surgeries, which are relatively short compared to the other specialisms. The turnover time for Ophthalmology was also significantly shorter than the other specialism, as can be seen in Figure 3.5. HAN also had a fewer variety of surgeries, while General Surgery, the specialism with the most types, had the highest overall turnover time. This could show that having only specific types of surgeries increases efficiency in the operating room and turnover time, as fewer changes

need to be performed in the operating room concerning for example the medical equipment. The results showed an average overtime of about 58 minutes on overall, however, the number of surgeries that had overtime was not that much. About 15% of the surgeries which were scheduled at the end of the day had overtime. However, this only included elective surgeries, as the emergency surgeries could be scheduled at any moment. Furthermore, HAN had again a significant lower overtime on overall compared to the other locations. This could be due to the smaller variation in surgery duration of the type of procedures performed in HAN, or due to a different method of scheduling. For example, the scheduler of HAN could have focussed more on decreasing overtime, while the others had other priorities. This was not information that was directly extracted from the results.

Another observation of the overtime was the difference between specialisms. For example, Neurosurgery had almost no overtime at location HO. An explanation for this is that some surgeries are known to deviate a lot, especially the more complex and longer surgeries. Knowing this, the schedulers could always have chosen to schedule those surgeries at the beginning of the day instead of the end. This would give these surgeries the time needed if there would be a delay, and the ability to schedule a simpler and shorter surgery at the end if the surgery would go faster than expected.

The undertime has the same problem as the turnover time. It is higher than actual operating room time that can be used for surgeries, as it does not include for example the cleaning of the operating room. Furthermore, this only included the elective surgeries, it could be that there were emergency surgeries scheduled after the last elective surgery. Hospitals commonly dedicate an operating room especially for emergency surgeries, which would mean that these surgeries would not influence the undertime as presented in the results. However, it could be that the emergency operating room was already used, making the other empty operating room more suitable for an emergency case.

Analyzing the operating room workflow based on the expert interviews, the ORPCs indicated the difficulties of creating an optimal schedule. The ORPCs use a more reactive method to change the schedule, where changes are made after for example a delay had occurred. This showed that situations occurred where reacting creates an unnecessary delay, where for example the next patient needs time to be able to arrive at the surgical department. If an ORPC anticipated the delay earlier, he/she could have called the patient before the end of the surgery. Using an estimation system could add value to this part, by giving the ORPCs the possibility to communicate with the patient earlier than currently possible. This would decrease the delay, which would decrease the turnover time. Instead of reacting to the events, using such a system would give them the possibility to anticipate changes. However, to be able to know the real added value of this system, an analysis needs to be made in a real setting where a comparison is made between the schedule with and without such a system.

Moreover, another addition of the estimation system is that it is automatic. As the surgeons' progress estimation is not considered as accurate by the ORPCs, an automatic system would be unbiased and estimate the duration only based on subjective events. Also, the scheduler would not need to disturb the medical staff in the operating room to get an estimate, as the system would provide this. This would give the ORPCs the possibility to check the remaining duration at any moment, instead of asking the surgeon once during surgery.

Another critical discussion point is based on the interviews performed and the results of those interviews. The interviews were performed with three operating room program coordinators (ORPCs), which is not a significantly large number of interviewees for making accurate conclusions. Furthermore, the hospital they work in could influence their perception of scheduling. However, the interviews gave merely a first indication of the current situation and how an estimation system would be useful in an operating room workflow.

### 4.3. Implementation in the OR workflow

Defining the implementation for an automatic intraoperative remaining surgery duration estimation system is key in creating a useful system. Based on the results of this research the following recommendations are made. First of all, on a system level, the main evaluation point of a system is the estimation accuracy. As indicated by the ORPCs, an estimation with a maximum error of 10 to 15 minutes would be valuable. Another measurement threshold that could be used would be an error of about 10% of the surgery duration, as this is relative to the duration instead of a fixed error. This would be about five minutes for a surgery of 50 min, and for the longer surgeries below the 15 min mark of the ORPCs. This needs, however, to be evaluated in a real-life setting, as this all depends on the scheduling method that would be applied in combination with an estimation system. Also, the systems estimation accuracy should be evaluated based on the error relative to the progress, as an estimation at the beginning is less important than in the middle. Currently, the methods described in literature are mostly interested in the accuracy overall. However, this research showed that the accuracy of Q3 could be more significant for the evaluation of the system than the other quarters.

Furthermore, another point that should be considered is who should have access to the estimation. Possibly, the estimation should not be communicated to the surgeons and the operating room staff during surgery. This could create unintentionally unnecessary pressure for the medical staff to finish faster. For this reason, the choice could be made that only the schedulers should have access to be able to change the schedule. Also, the ORPCs indicated the need to continuously check the estimation system to be able to follow the progress, which would not be possible in the current workflow. Instead, the system could give a notification when the surgery duration would deviate from the schedule based on a range. This would make sure the schedulers are alerted, without the need of continuously checking the system.

### 4.4. Future research

The main sources for the estimation of the remaining surgery duration are the phases of the surgery. The methods described in this research are highly dependent on the phases, as these indicate the progress. As described by Meij [20], a neural network could be used to predict the phases of the system. However, to be able to do so, the system needs training data which was created manually. This is time consuming, as this would have to be done for every type of surgery. As described in section 2.1, each type of surgery had different phases. Another possibility could be to create generic phases, which a system could predict automatically. Instead of creating phases that are specific for a type of surgery, a prediction could be made for phases without labeling the phases. By using a large number of videos as training data, an unsupervised classification system could create generic phases for the training data. Using a method such as the novel method, the system would automatically find similar videos to estimate the remaining surgery duration, without knowing which type of surgery it is. The similarity measurement could use the type of surgery as an estimation factor, but it is not necessarily needed for the phases. This would solve the issue for non-sequential phases, such as the morcellation and transvaginal retrieval in the Total Laparoscopic Hysterectomy dataset, as these would be automatically grouped in a different group of videos. This would give the possibility to increase the data drastically, as any surgery done in the past could be automatically added. The definition of surgeries and calling the surgery for example a Laparoscopic Cholecystectomy surgery was created by experts in this field but could be grouped for a machine more efficiently in an entirely different way, such as possible a Simple Laparoscopic Cholecystectomy and another group called Complex Laparoscopic Cholecystectomy. Especially for the estimation of the remaining surgery duration, it would be useful to find similar surgeries not necessarily based on

a name but the similarity of the duration. However, the downfall of such a system is that it must use a dynamic method, such as the novel method, to estimate the remaining surgery duration. A fixed method, such as the phase-inferred method, would not be possible as the method is not able to distinguish between different types of surgeries and would eventually use all the videos to estimate the remaining duration. Furthermore, to the authors' knowledge, no research has been done in automatic creating generic phases. Research will have to be done to the type of method to create these phases, and how to define for example the number of phases.

Also, to be able to use methods such as the novel method as described in this research, defining the right similarity measurement and the features for this measurement are key for an optimal system. Therefore, more extensive research has to be done to which features to be used and what type of similarity measurement is optimal. The current similarity measurement is the Dynamic Time Warping method combined with the Euclidian distance, a method frequently used in time-series forecasting. However, little research had been done to the right similarity measurement that would be useful for this field. As surgery progress is a kind of time-series, DTW is a right approach, but other methods could be more optimal for this system. Also, the features used were selected as these features were available. It could be beneficial to do more extensive research on which features would identify similar surgeries as, for example, preoperative data such as BMI and ASA score could be useful for the estimation of the remaining surgery duration.

Furthermore, the current scheduling process possibly needs to change to make an automatic estimation more beneficial. For example, the current schedule is changed manually during the day. However, if an automatic estimation system would have the right performance, an automatic scheduling system could be created so that there would be no need for manual changes (or only small necessary changes).

## 4.5. Conclusion

This research was focussed on the creation and evaluation of an automatic intraoperative remaining surgery duration estimation system based on surgical phases to improve the operating room workflow. It showed that such a system was possible to be created with an accuracy of about 10 minutes; an accuracy that would be acceptable based on expert interviews. It demonstrated different methods that could create an estimation system, and how these methods performed. Using similar videos in combination with phase statistics could outperform known state-of-the-art methods. The analysis and interviews about the operating room workflow gave an impression on where such a system could benefit the operating room workflow. For example, continuous surgery progress analysis could decrease the turnover time as this system would allow anticipating surgery duration diversion in the schedule. Also, the surgeons' estimation bias could be avoided using such a system. Overall, this research showed the feasibility and usefulness of an intraoperative remaining surgery duration system based on surgical phases for the operating room workflow. The next step would be to understand how to implement such a system in the operating room workflow. Who should be allowed to access the estimation, and how do you communicate this estimation to the user? An automatic intraoperative remaining surgery duration estimation system showed to be promising for improving the operating room workflow, however, which changes are needed in the OR workflow to be able to use such a system?





The background features a complex, light blue wireframe structure on the left side, resembling a stylized letter 'B' or a similar geometric form. This structure is composed of numerous interconnected lines and small dots. Below the title, a horizontal line is decorated with several clusters of interconnected nodes and lines, creating a network-like appearance. The overall aesthetic is clean, modern, and technical.

# BIBLIOGRAPHY

- [1] B. Cardoen, E. Demeulemeester, and J. Beliën, "Operating room planning and scheduling: A literature review," *European journal of operational research*, vol. 201, no. 3, pp. 921–932, 2010.
- [2] A. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier, "Staff scheduling and rostering: A review of applications, methods and models," *European Journal of Operational Research*, vol. 153, no. 1, pp. 3 – 27, 2004.
- [3] B. Akbarzadeh, G. Moslehi, M. Reisi-Nafchi, and B. Maenhout, "The re-planning and scheduling of surgical cases in the operating room department after block release time with resource rescheduling," *European Journal of Operational Research*, vol. 278, no. 2, pp. 596–614, 2019.
- [4] M. Yoshikawa, K. Kaneko, T. Yamanouchi, and M. Watanabe, "A constraint-based high school scheduling system," *IEEE Expert*, vol. 11, no. 1, pp. 63–72, 1996.
- [5] J. Appleby, D. Blake, and E. Newman, "Techniques for producing school timetables on a computer and their application to other scheduling problems," *The Computer Journal*, vol. 3, no. 4, pp. 237–245, 1961.
- [6] E. K. Baker, "An exact algorithm for the time-constrained traveling salesman problem," *Operations Research*, vol. 31, no. 5, pp. 938–945, 1983.
- [7] A. Wu, D. E. Rinewalt, R. W. Lekowski Jr, and R. D. Urman, "Use of historical surgical times to predict duration of primary aortic valve replacement," *Journal of cardiothoracic and vascular anesthesia*, vol. 31, no. 3, pp. 810–815, 2017.
- [8] E. R. Edelman, S. M. J. van Kuijk, A. E. W. Hamaekers, M. J. M. de Korte, G. G. van Merode, and W. F. F. A. Buhre, "Improving the prediction of total surgical procedure time using linear regression modeling," *Frontiers in Medicine*, vol. 4, p. 85, 2017. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fmed.2017.00085>
- [9] A. Wu, M. J. Weaver, M. M. Heng, and R. D. Urman, "Predictive model of surgical time for revision total hip arthroplasty," *The Journal of arthroplasty*, vol. 32, no. 7, pp. 2214–2218, 2017.
- [10] C. Thiels, D. Yu, A. Abdelrahman, E. Habermann, S. Hallbeck, K. Pasupathy, and J. Bingener, "The use of patient factors to improve the prediction of operative duration using laparoscopic cholecystectomy," *Surgical Endoscopy*, vol. 31, 01 2017.
- [11] F. Dexter, A. Macario, R. H. Epstein, and J. Ledolter, "Validity and usefulness of a method to monitor surgical services' average bias in scheduled case durations," *Canadian Journal of Anesthesia/Journal canadien danesthésie*, vol. 52, no. 9, p. 935–939, 2005.
- [12] B. M. Gillespie, W. Chaboyer, and N. Fairweather, "Interruptions and miscommunications in surgery: an observational study," *AORN journal*, vol. 95, no. 5, pp. 576–590, 2012.

- [13] A. N. Healey, N. Sevdalis, and C. A. Vincent, "Measuring intra-operative interference from distraction and interruption observed in the operating theatre," *Ergonomics*, vol. 49, no. 5-6, pp. 589–604, 2006, PMID: 16717011.
- [14] S. Franke, J. Meixensberger, and T. Neumuth, "Intervention time prediction from surgical low-level tasks," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 152–159, 2013.
- [15] F. C. Meeuwsen, F. van Luyn, M. D. Blikkendaal, F. W. Jansen, and J. J. van den Dobbelsteen, "Surgical phase modelling in minimal invasive surgery," *Surgical Endoscopy*, vol. 33, no. 5, pp. 1426–1432, 2019.
- [16] N. Spangenberg, M. Wilke, and B. Franczyk, "A Big Data architecture for intrasurgical remaining time predictions," in *Procedia Computer Science*, vol. 113. Elsevier B.V., 2017, pp. 310–317.
- [17] A. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "RSDNet: Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1069–1078, 2019.
- [18] A. C. Guédon, M. Paalvast, F. C. Meeuwsen, D. M. Tax, A. P. van Dijke, L. S. Wauben, M. van der Elst, J. Dankelman, and J. J. van den Dobbelsteen, "It is Time to Prepare the Next patient' Real-Time Prediction of Procedure Duration in Laparoscopic Cholecystectomies," *Journal of Medical Systems*, vol. 40, no. 12, 2016.
- [19] S. Bodenstedt, D. Rivoir, A. Jenke, M. Wagner, M. Breucha, B. Müller-Stich, S. T. Mees, J. Weitz, and S. Speidel, "Active learning using deep Bayesian networks for surgical workflow analysis," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 6, pp. 1079–1087, 2019.
- [20] S. Meij, "The applicability of deep learning to detect the progress of laparoscopic surgery using video recordings," 2019. [Online]. Available: <http://resolver.tudelft.nl/uuid:28f96ac7-f5ae-43af-9ba0-14832af5c103>
- [21] W. McKinney, *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc., 2012.
- [22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [23] R. Nakamura, T. Aizawa, Y. Muragaki, T. Maruyama, and H. Iseki, "Method for End Time Prediction of Brain Tumor Resections Using Analysis of Surgical Navigation Information and Tumor Size Characteristics," *Tech. Rep.*, 2012. [Online]. Available: [www.springerlink.com](http://www.springerlink.com)
- [24] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2012, vol. 821.
- [25] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *Journal of clinical epidemiology*, vol. 49, no. 11, pp. 1225–1231, 1996.

- [26] H. Hassan, A. Negm, M. Zahran, and O. Saavedra, "Assessment of artificial neural network for bathymetry estimation using high resolution satellite imagery in shallow lakes: Case study el burullus lake." *International Water Technology Journal*, vol. 5, 12 2015.
- [27] G. Hackeling, *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd, 2017.
- [28] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.
- [29] T. Koskela, M. Lehtokangas, J. Saarinen, and K. Kaski, "Time series prediction with multilayer perceptron, fir and elman neural networks," in *Proceedings of the World Congress on Neural Networks*. Citeseer, 1996, pp. 491–496.
- [30] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [31] S. D. Jadhav and H. Channe, "Comparative study of k-nn, naive bayes and decision tree classification techniques," *International Journal of Science and Research (IJSR)*, vol. 5, no. 1, pp. 1842–1845, 2016.
- [32] A. Singh, A. Yadav, and A. Rana, "K-means with three different distance metrics," *International Journal of Computer Applications*, vol. 67, no. 10, 2013.
- [33] G. Qian, S. Sural, Y. Gu, and S. Pramanik, "Similarity between euclidean and cosine angle distance for nearest neighbor queries," in *Proceedings of the 2004 ACM symposium on Applied computing*, 2004, pp. 1232–1237.
- [34] K. Chomboon, P. Chujai, P. Teerarassamee, K. Kerdprasop, and N. Kerdprasop, "An empirical study of distance metrics for k-nearest neighbor algorithm," in *Proceedings of the 3rd international conference on industrial application engineering*, 2015, pp. 1–6.
- [35] W. A. Chaovalitwongse, Y.-J. Fan, and R. C. Sachdeo, "On the time series  $k$ -nearest neighbor classification of abnormal brain activity," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 37, no. 6, pp. 1005–1016, 2007.
- [36] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.
- [37] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.
- [38] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and information systems*, vol. 7, no. 3, pp. 358–386, 2005.
- [39] A. Vehtari, A. Gelman, and J. Gabry, "Practical bayesian model evaluation using leave-one-out cross-validation and waic," *Statistics and computing*, vol. 27, no. 5, pp. 1413– 1432, 2017.
- [40] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.



# APPENDIX A INTERVIEW RESULTS

## Interview 1

- **Interviewee:** ORPC 1
- **Location:** Haarlem-South (HAZ)
- **Date:** 11th of June, 2020

***Q1. How do you make sure that the schedule is correct during the day? What do you change?***

The operating room scheduling system continuously changes itself, when for example the surgery started later than expected or has a delay. In this case, the system pushes the next surgery so that the start time of the next surgeries is later. Based on the changing schedule, I try to rearrange the schedule so that all the surgeries can still be performed.

***Q2. Which information do you need to during the day to be able to change the schedule properly?***

Everything. I try to communicate with everybody to know their status. And check regularly the current progress by sight.

***Q3. How do you retrieve information about the progress of current surgeries?***

I continuously communicate with the OR staff the current progress and look into the operating room. Based on my own experience, I can see the current progress of the surgery and anticipate the duration. Furthermore, I communicate with everybody to know their current non-surgical task, to be able to estimate the progress. If someone is getting coffee, that means a surgery is over. I also get an estimation of the surgeon.

***Q4. Based on your experience, is the remaining surgery duration estimation of the surgeon during the surgery correct?***

I know the people. With some surgeons I do trust their estimation, but with others I know that they sometimes underestimate it. So, it is not always correct, but I can judge the progress myself based on the current events in the operating room.

***Q5. Are there types of surgeries where the surgery duration is more difficult to predict than others?***

Of course, but those surgeries are labeled in the system as longer. I always call the surgeons of the next day to ask if it will be longer or shorter than on average, and preventively change the schedule accordingly.

***Q6. Do you change the schedule during a surgery? If so, in which part of the surgery? (Q1, Q2, Q3, or Q4)***

Never at the beginning. Continuously, mostly at the end. I ask the surgeon only once for an estimation, as this is creating disturbance in the operating room. It would be useless to ask it at the beginning and cause unnecessary frustrations. But by looking through the window, I can see the current situation and that way know the progress.

**Q7. In your experience, how much is the turnover time?**

On average about 30 minutes between each surgery. But it deviates a lot.

**Q8. Is this necessary/needed, or too much/too little?**

Yes, it all depends on the patient. He/she needs to change between rooms and that could take more time. Currently it takes even longer due to Covid-19, because of the change of sterile gloves and masks.

**Q9. Is there a high occurrence of overtime for the operating room staff? (Daily, once a week,.. etc)**

Every day, but it is not a lot of overtime.

**Q10. What is the reason for overtime?**

A lot of reasons. For example, surgeries are longer than expected.

**Q11. Is there a high occurrence of undertime for the operating rooms? (Daily, once a week,.. etc)**

Never, if it could be empty at the end of the day, I will schedule another surgery.

**Q12. What is the reason for undertime?**

-

**Q13. Would you want to use an automatic surgery duration estimation system, and why (not)?**

No, such a system won't work. There are too many factors to be able to estimate the duration. I need to know everything of every person, and a system is not able to do that.

**Q14. How would such a system help your daily work?**

-

**Q15. If you would use an automatic remaining surgery duration estimation system, what should the accuracy be to be acceptable?**

-

**Q16. When during the surgery do you need the estimation? (Q1, Q2, Q3, Q4)**

-



## Interview 2

- **Interviewee:** ORPC 2
- **Location:** Hoofddorp (HO)
- **Date:** 11th of June, 2020

**Q1. How do you make sure that the schedule is correct during the day? What do you change?**

We continuously check the progress of the operating rooms. There are three rooms allowed to have overtime, two for the elective surgeries and one for the emergency surgeries. For example, if there is a surgery with a delay, we try to rearrange for example the next surgery in another room.

**Q2. Which information do you need to during the day to be able to change the schedule properly?**

The progress of the patient, so where the patient is. Is he/she in the holding area, or still on his/her way? And the progress of the surgeries.

**Q3. How do you retrieve information about the progress of current surgeries?**

We use the timestamps to check the progress, such as start incision, start closing etc. We can use these timestamps to understand where the delay is, and this way change the schedule accordingly. Furthermore, we continuously check the surgeries by looking into the operating room and know the progress. And if there is a complication, we get noticed by the OR staff.

**Q4. Based on your experience, is the remaining surgery duration estimation of the surgeon during the surgery correct?**

It is just an indication; it is hard for them to estimate. It is just an indication on a certain moment. But a complication could be just after that.

**Q5. Are there types of surgeries where the surgery duration is more difficult to predict than others?**

Yes, those are categorized as a different surgery.

**Q6. Do you change the schedule during a surgery? If so, in which part of the surgery? (Q1, Q2, Q3, or Q4)**

Continuously, based on the delays in the system. It is based on the surgery progress, but when something happens such as a delay or a surgery that is finished earlier.

**Q7. In your experience, how much is the turnover time?**

On average 30 min.

**Q8. Is this necessary/needed, or too much/too little?**

It is necessary, it is not possible to change the patients directly. There are multiple factors that could influence the turnover time, such as more time needed for the cleaning, or a delay in change of surgeon or anesthesiologist. Also, if a surgery is finished earlier than expected, the next patients need to be informed and could possibly be not at the hospital yet. In that situation we call, and try to ask the patient to come earlier, but we are fully dependent on them.

**Q9. Is there a high occurrence of overtime for the operating room staff? (Daily, once a week,.. etc)**

Not a lot after 16.30 no. On average once a week.

**Q10. What is the reason for overtime?**

Could be anything. For example, a surgery was delayed because of an instrument that was dropped.

**Q11. Is there a high occurrence of undertime for the operating rooms? (Daily, once a week,.. etc)**

Almost never, overtime more frequently.

**Q12. What is the reason for undertime?**

A cancellation of surgery because for example the patient needed something or instrument that is not available.

**Q13. Would you want to use an automatic surgery duration estimation system, and why (not)?**

It sounds useful, but it does not show the progress of the total workflow. This also influences the schedule. Also, I will have to continuously check the system, which would take a lot of time next to my current work. I would not think that it would add a lot of information.

**Q14. How would such a system help your daily work?**

If I continuously check the system, it could be useful. But that is not possible next to my current tasks.

**Q15. If you would use an automatic remaining surgery duration estimation system, what should the accuracy be to be acceptable?**

10 min.

**Q16. When during the surgery do you need the estimation? (Q1, Q2, Q3, Q4)**

Q3

## Interview 3

- **Interviewee:** ORPC 3
- **Location:** Hoofddorp (HO)
- **Date:** 11th of June, 2020

### ***Q1. How do you make sure that the schedule is correct during the day? What do you change?***

We use the Epic System to know the average durations for each surgery. Furthermore, we check if the scheduled duration of each surgery is right. For example, If Surgeon A is going to perform a cholecystectomy and the scheduled surgery duration is 2 hours, but we know he is always faster than average, we change the schedule accordingly. We check the reliability of the schedule. Also, ambulatory is the department that schedules the surgeries. Sometimes it is a different surgery than an average surgery of that type, which is added as a note in the surgery request. Based on that we can change the schedule to make sure it is kind of correct. During the day it is difficult, there is no one way. We react on the events happening, such as delay or faster surgery. And if there is a large turnover, we ask the staff what is happening and the reason of this turnover. But it is all reacting, no predicting/preventive changes in the schedules. There are too much factors to be able to have a good prediction.

### ***Q2. Which information do you need to during the day to be able to change the schedule properly?***

The notes of ambulatory, so patient/surgery specific information. And if there is a complication during a surgery. And the current progress.

### ***Q3. How do you retrieve information about the progress of current surgeries?***

If there is for example a change of surgery, or complication, I get noticed by the OR staff. And I can see if the surgery is currently taking more time than scheduled.

### ***Q4. Based on your experience, is the remaining surgery duration estimation of the surgeon during the surgery correct?***

Mostly, but I use my personal experience. Some surgeons are correct, others for example always underestimate.

### ***Q5. Are there types of surgeries where the surgery duration is more difficult to predict than others?***

Yes, especially the longer surgeries than average. That can deviate a lot. It could easily be an hour longer than estimated or finish really fast. This influences the schedule significantly, while the shorter surgeries are more standard.

### ***Q6. Do you change the schedule during a surgery? If so, in which part of the surgery? (Q1, Q2, Q3, or Q4)***

Continuously, but it is based on for example a delay. So, when I see the surgery is still in progress, longer than scheduled, than I can change the next schedule.

### ***Q7. In your experience, how much is the turnover time?***

It differs a lot, but about 20 minutes. But It could easily be an hour.

**Q8. Is this necessary/needed, or too much/too little?**

There are multiple reasons for a turnover time. Sometimes the people are faster than normal. For example, a patient needs to be called. Sometimes it takes one person more time to do so than the other. Or what also happens sometimes is that we forget to call the next patient. So, we are waiting and remember that we did not call yet. Furthermore, the department that needs to bring the patient could be slower or faster. That could differ from 15 min to 45 min. Or the holding is busy, which could create a longer turnover time. So, it is mostly because we need to wait on the patient to be ready.

**Q9. Is there a high occurrence of overtime for the operating room staff? (Daily, once a week,.. etc)**

Not that much.

**Q10. What is the reason for overtime?**

Most of the time due to a longer turnover time.

**Q11. Is there a high occurrence of undertime for the operating rooms? (Daily, once a week,.. etc)**

Not that much, maybe a little bit, but not enough time for a new surgery.

**Q12. What is the reason for undertime?**

-

**Q13. Would you want to use an automatic surgery duration estimation system, and why (not)?**

Yes, it would be useful, but I do not know how. Maybe it would make it possible to change the schedule earlier, but I do not think it would change the total schedule at the end.

**Q14. How would such a system help your daily work?**

I don't know.

**Q15. If you would use an automatic remaining surgery duration estimation system, what should the accuracy be to be acceptable?**

About 10 to 15 minutes is acceptable.

**Q16. When during the surgery do you need the estimation? (Q1, Q2, Q3, Q4)**

Q3, but I would prefer it more at the beginning of Q3.



