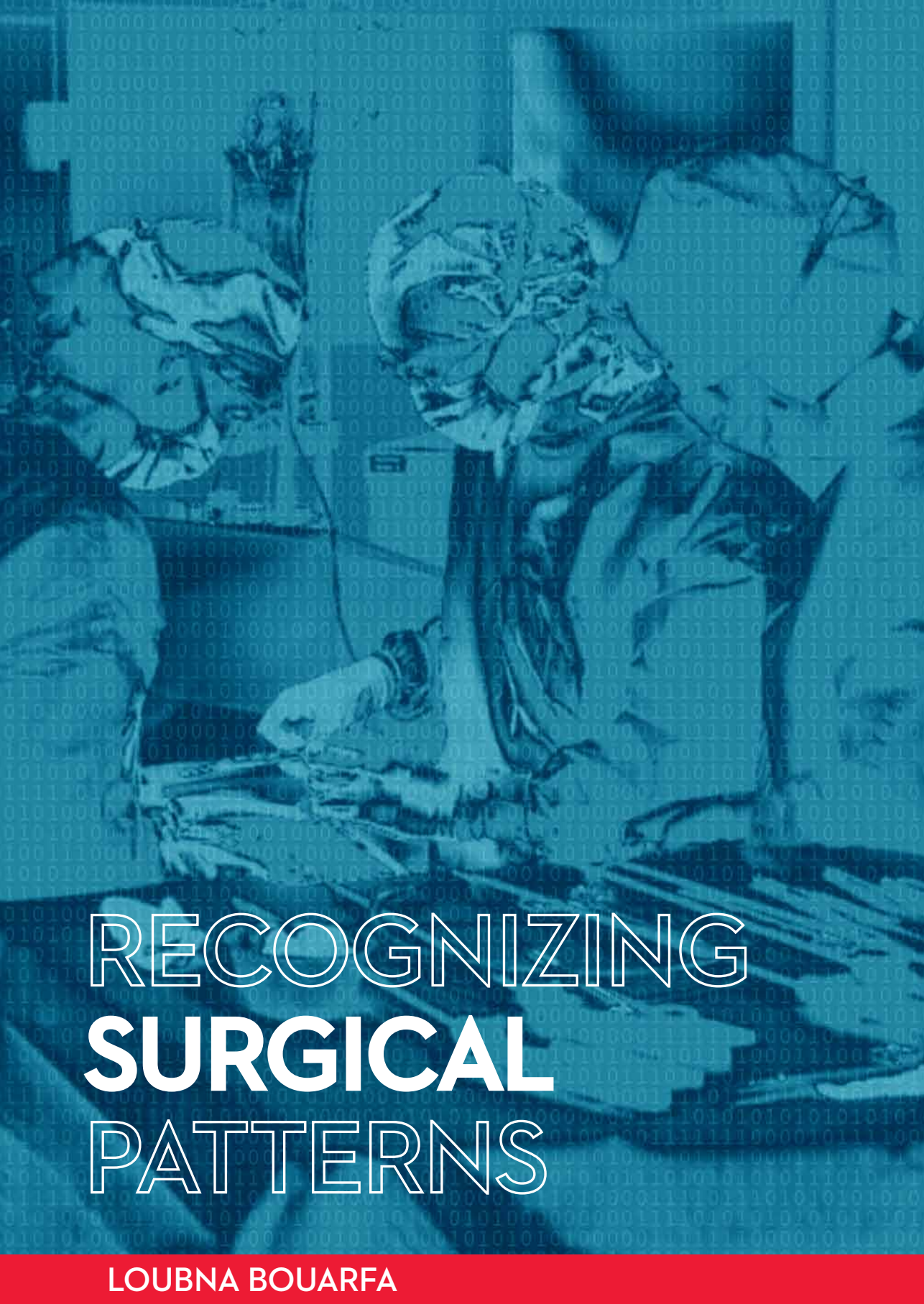


RECOGNIZING SURGICAL PATTERNS

LOUBNA BOUARFA

Recognizing surgical patterns

Loubna Bouarfa

Recognizing surgical patterns

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op 30 mei 2012 om 12:30
uur
door

Loubna BOUARFA

Ingenieur in Media & Knowledge Engineering
geboren te Meknès, Marokko.

Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. J. Dankelman
Prof. dr. ir. P.P Jonker

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. J. Dankelman	Technische Universiteit Delft, promotor
Prof. dr. ir. P.P. Jonker	Technische Universiteit Delft, promotor
Prof. dr. F.W. Jansen	Leiden Universitair Medisch Centrum
Prof. dr. M. Neerincx	Technische Universiteit Delft
Prof. dr. J. Klein	Erasmus Universiteit Rotterdam
Prof. dr. N. Navab	Technischen Universität München
Dr. T. Weijters	Technische Universiteit Eindhoven
Prof. dr. ir. C.A. Grimbergen	Academisch Medisch Centrum, reservelid

This research and the production of this thesis has been financially supported by the Dutch grant organization STW, project number: 07320.

ISBN:978-94-6169-251-1

Author email: loubna.bouarfa@gmail.com

Contents in brief

Contents

Chapter 1

Introduction

1.1 Motivation

The operating room (OR) is the most costly environment in healthcare [Jea10; Hoz05], it requires expensive labour resources (surgeon, anaesthetists, nurses, etc.), high-priced equipment and high daily maintenance.

In the Netherlands, each year more than 1700 patients die from preventable surgical errors [DB07]. Medical errors are commonly referred to as Adverse Events (AEs). Surgical AEs account for one-half to three-quarters of all AEs in healthcare [Gri08]. The Dutch Patient Safety Research Program showed that AEs affect 5.7% patients in Dutch hospitals and result in permanent disabilities, morbidities and even mortalities. Forty per cent of these surgical AEs were judged preventable by following established clinical practices [Zeg07]. Using an aggregated estimated cost of 1100 euro per hour for an OR, the cost of preventable, surgical AEs in the Netherlands is estimated at 161 million euro in 2004 alone [Hoo09]. Therefore preventing AEs can save costs as well as improve patient safety.

Another aspect of OR management is the utilization rate of the OR. Both under- and over-utilization of the operating room represent unnecessary (and unstable) costs for hospitals [Car10b]. A benchmark study showed that the average delay in surgery start times ranges from 25 to 103 minutes [Doe09]. This is shown to be caused by the failure of commonly used planning tools which do not account for the unpredictable time duration of surgery. Capacity problems can also be reduced by avoiding unexpected events, like longer procedure times because of complications [Hoo09]. In the Netherlands those lost hours add up to 2150 hours, equivalent to 2.3 million euro per year.

In production industry, standardization of the production processes is deployed to both reduce errors and improve efficiency by allowing simpler planning. Surgery is characterized by a peri-operative pipeline of pre-, intra- and post-operative processes that can be instantiated with a workflow model. To both reduce errors

and improve efficiency, the workflow in the peri-operative pipeline should be designed and planned as effectively as possible in terms of flow of patients and allocation of scarce resources such as operating rooms, instruments and personnel [Neu06b; Neu08]. It is well known for pipeline systems, such as in industrial production systems and traffic systems, that reduction in fluctuations within the system parts has a positive effect on the throughput of the entire pipeline. The peri-operative pipeline varies from country to country and even from hospital to hospital (e.g. the number of ORs vs. PACUs may vary) and their surgical workflow may vary as well. Moreover, the amount and quality of the data used for planning varies. However, for all implementations the rule is that fluctuations should be suppressed in all sub-processes of the pipeline to optimise the throughput. If this goal is not feasible then at least the fluctuations should be made predictable, so a planning system that covers the entire workflow of the peri-operative pipeline can efficiently cope with the effects of those fluctuations. A first step is to standardize the surgical workflow as is common in production engineering, air traffic control and the military.

Standardization of surgery is, however, a very hard challenge because surgeons vary in their experience in and ability to perform a surgical technique, there are individual preferences in performing the procedure, and technical modifications may occur as the procedure evolves [Wol07]. Furthermore, each patient has -to a degree- a different anatomy and surgeons are trained to adapt their methods to match those differences. When many surgeons, with different skill levels, perform different procedures on patients, each with a different anatomy, they may agree on the standardization of most critical aspects of the procedures, but it is almost impossible to reach consensus on all aspects.

1.2 Focus on laparoscopic surgery

The work in this thesis focuses on laparoscopic surgery as it is used in a comparable setup for different surgical specialities. Figure 1.1 illustrates the general setup of laparoscopic surgery. The procedure is performed through small incisions in the abdominal wall to gain access to the internal anatomy of the patient's body. To be able to work via these small incisions, specialised thin instruments and a camera are used, as is illustrated in Figure 1.1. The main benefits of laparoscopic surgery, when compared to traditional open surgery are: reduced inconvenience to patients, reduced patient's trauma, shortened hospitalisation, improved diagnostic accuracy and improved therapeutic outcome. The downside to these benefits is that laparoscopic surgery requires a high degree of manual dexterity from the surgeon (or operator) as the instrument controls are more complex when compared to open surgery.

Figure 1.2 illustrates a schematic diagram of different applications of laparoscopic techniques in surgery. The most common laparoscopic procedure performed is laparoscopic cholecystectomy, which is illustrated in Figure 1.2 (a). Laparo-

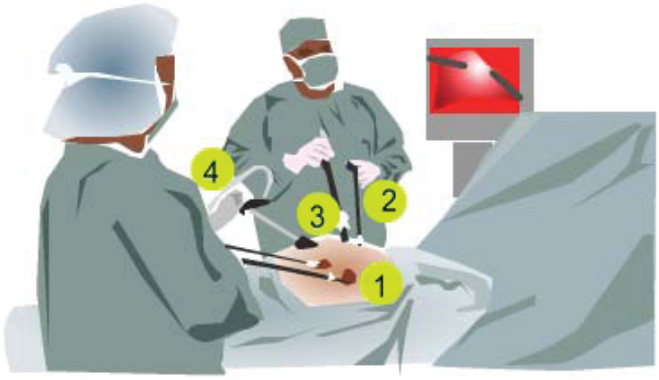


Figure 1.1: *A schematic diagram from [Lo07] illustrating the basic setup of laparoscopic surgery: small incisions are made on the abdomen and the laparoscopic instruments are inserted using trocars. The surgeon performs the surgery by watching the feed from the laparoscopic camera projected onto a video screen. (1) Laparoscopic incision, (2) Laparoscopic instruments, (3) trocars, and (4) Laparoscopic camera.(Used by permission from [Lo07])*

scopy was first applied to gynaecology in 1960. The technique has now been applied to a variety of specialties [Lo07]. In gynaecology the technique is applied to a variety of procedures such as laparoscopic hysterectomy, laparoscopic supracervical hysterectomy, and laparoscopic vault suspension. Laparoscopic hysterectomy is illustrated in Figure1.2 (b). In urology, laparoscopy is used for the treatment of kidney tumours and nephrectomies as is depicted in Figure1.2 (c). In most of the laparoscopic procedures an organ is removed.

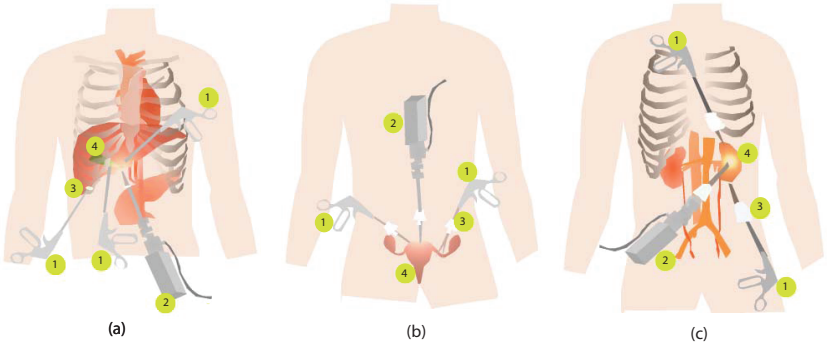


Figure 1.2: *A schematic diagram illustrating the setup of: (a) Laparoscopic cholecystectomy, (b) Laparoscopic hysterectomy, (c) Laparoscopic nephrectomy. (1) Laparoscopic instruments, (2) Laparoscopic camera, (3) Abdominal incisions and trocars, and (4) The organ to be removed.(Used by permission from [Lo07])*

1.3 Opportunities in laparoscopic surgery

Generally, in the entire peri-operative process (pre-, intra- and post-operative) data is gathered that can be used to minimize AEs and boost efficiency by enhancing the predictability and standardization of laparoscopic surgery.

Before the start of surgery, **pre-operative data** is collected; including clinical history of the patients, current and past medication, measured vital signs, laboratory data, radiology examination, nursing records and operation records. Moreover, pre-operative imaging is occasionally used to plan the surgical strategy for localization and size assessment purposes.

During surgery, the growing availability of measurement devices used in the OR enables the collection of a large volume of **intra-operative data** about the course of surgery and the state of the patient during surgery. All procedures combined produce many hours of endoscopic video each day. Moreover, the patient's blood pressure, heart rate, cardiac rhythm, expired CO_2 and body temperature is routinely measured. After surgery, patients are transferred to the Post Anaesthesia Care Unit (PACU) or the Intensive Care Unit (ICU) to recover. Here **post-operative data** is collected. Sophisticated monitoring equipment performs measurements of multiple physiological parameters on a high frequency. Those measurements require timely -and context sensitive- analysis in order to sustain effective decision support [Sam06]. Because so much data could be available, it is important to identify those parameters most important to predicting adverse events and increasing the efficiency of the OR and its related resources.

1.4 Objective

The goal of this thesis is to show how off- and on-line acquired peri-operative data can be analysed with pattern recognition techniques to reduce adverse events and improve efficiency of surgical procedures. The most important questions addressed are:

- How to use available prior knowledge to improve operative safety?
- How to measure surgical workflows (i.e. standards) regardless of the level of variance of its execution by different surgeons and for different patients?
- How to automatically acquire surgical workflow data?
- How to automatically detect possible adverse events during surgical interventions?
- How to use peri-operative data to support peri-operative planning?

The work in this thesis focuses on processing pre-, inter- and post-operative data using pattern recognition techniques to predict safety and efficiency parameters in surgery. To the best of the author's knowledge the use of pattern

recognition techniques with surgical data is limited in the literature to imaging applications. Surgical data is currently processed using common clinical statistical tools. We expect, however, that future studies in surgery will employ advanced pattern recognition tools in processing surgical data to improve safety and efficiency of surgeries. The research presented in this thesis is a first step towards that direction.

1.5 Contributions & outline

The work in this thesis is focused on logging and predicting events from data available before or in the OR. Pattern recognition (PR) techniques are used as they can accommodate less strict data collection (i.e. especially when compared to Randomized controlled trials). PR techniques provide mathematical tools to either manage shortcomings of surgical data (e.g. bias, limited samples) and can provide the best possible results with the given data. The contributions of this thesis are divided in the following chapters:

- **Chapter 2** Explores the limitations of using Randomized Controlled Trials in surgery and the feasibility of using pattern recognition to measure safety, effectiveness and efficiency of surgical treatments. This work was presented in [Bou12b].

- **Chapter 3** Pre-operative patient data is used in this chapter to train a classifier and to identify the interesting features for predicting intra-operative complexity. This work was presented in [Bou11c].

- **Chapter 4** Presents a framework for recognizing high-level surgical tasks from low-level sensors in the operating room. We further show how to use this framework to detect surgical activities from instrument signals using laparoscopic video as input. This work was presented in [Bou11b].

- **Chapter 5** Instrument signals used in the previous chapter were manually annotated from the laparoscopic video. This chapter presents a tracking system to detect and track instruments in laparoscopic video using biocompatible colour markers. This work was presented in [Bou11a].

- **Chapter 6** Presents a new approach for deriving surgical consensus from running surgeries. The derived consensus is proven to conform the main steps of laparoscopic cholecystectomy as defined in best practices. The paper also shows how outliers can be detected from process logs using the derived consensus [Bou12a].

- **Chapter 7** Peri-operative data is used in this chapter to predict the length of stay of patients in the Post Anaesthesia Care Unit (PACU). This work was presented in [Bou12c].

- **Chapter 8** Discussions, conclusions and future research directions are given here.

1.6 Publication List

1.6.1 List of Journal papers

The following papers, written by the author of this thesis, have been published or are currently under peer review:

1. L.Bouarfa, D. Tax and J. Dankelman, "Pattern recognition: a new perspective for evidence based surgery", Submitted article (Reference [Bou12b])

2. L. Bouarfa, A. Schneider, H. Feussner, N. Navab, H. U. Lemke, P.P. Jonker and J. Dankelman, "Prediction of intraoperative complexity from preoperative patient data for laparoscopic cholecystectomy", Journal of Artificial Intelligence in Medicine, vol. 52, pp.169-176, 2011 (Reference [Bou11c])

3. L. Bouarfa,P.P. Jonker, and J. Dankelman, "Discovery of high-level tasks in the operating room", Journal of Biomedical Informatics, vol.44, pp. 455-462, 2011 (Reference [Bou11b])

4. L. Bouarfa and J. Dankelman, "Measuring consensus and detecting outliers from surgical process logs", Submitted article (Reference [Bou12a])

5. L. Bouarfa, O. Akman, A. Schneider, P.P. Jonker and J. Dankelman, "In-vivo real-time tracking of surgical instruments in endoscopic video", Minimally Invasive Therapy & Allied Technologies, vol. 0, pp.0-6, 2011 (Reference [Bou11a])

6. L. Bouarfa, D. Tax, J.M. Ehrenfeld, B. Rothman and J. Dankelman, "Length of Stay in the Post Anaesthesia Care Unit - Can it be estimated?", Submitted article (Reference [Bou12c])

1.6.2 List of conference papers and abstracts

1. L. Bouarfa, L.P.S. Stassen, P.P. Jonker and J. Dankelman "Discovery of surgical high-level activities in the Operating Room", The Sixteenth annual conference of the Advanced School for Computing and Imaging (ASCI), Veldhoven, The Netherlands, November 1-3 2010

2. L. Bouarfa, O. Akman, A. Schneider, P.P. Jonker and J. Dankelman "Rapid detection of multiple instruments in endoscopic video", 22th International Conference of the Society for Medical Innovation and Technology (SMIT), Trondheim, Norway, September 2-4 2010

3. L. Bouarfa, P.P. Jonker and J. Dankelman "In-Vivo Measuring Surgical Workflow Activities in the OR", 7th international conference on methods and techniques in behavioural research, Eindhoven, The Netherlands, August 24-27 2010

4. L. Bouarfa and J. Dankelman "From Low-level sensors to high-level intelligence in the operating room", 18th International Congress of the European Association of Endoscopic Surgery (EAES), Geneva, Switzerland, June 16-19 2010

5. L. Bouarfa and J. Dankelman "Automatic surgical phase identification in the Operating Room", 21th International Conference of the Society for Medical Innovation and Technology (SMIT), Sinaia, Romania, October 7-9 2009

6. L. Bouarfa, P.P. Jonker and J. Dankelman "Surgical context discovery by monitoring low-level activities in the OR", MICCAI London, 1st Workshop on Modelling and Monitoring of Computer Assisted Interventions (M2CAI), September 20-24 2009

7. L. Bouarfa and J. Dankelman "High-level task modelling for the operating room", 17th International Congress of the European Association for Endoscopic Surgery (EAES), Prague, Czech Republic, June 17-20 2009

8. L.Bouarfa and J. Dankelman "Non-intrusive Optimization of Workflow in the Operating Room", 20th International Conference of Society for Medical Innovation and Technology (SMIT), Vienna, Austria, August 28-30 2008

Chapter 2

Pattern recognition: a new perspective for evidence based surgery

L.Bouarfa, D.M.J.Tax, J.Dankelman
Submitted article
under the title "*New perspectives for evidence based surgery: using pattern recognition to predicting outcome*"

abstract

Evidence-based medicine aims to utilize the best available evidence to support clinical decision making. To obtain the strongest evidence, usually Randomized Controlled Trials (RCTs) are used to measure the safety and efficacy parameters of new treatments. However, RCTs have many limitations when applied to surgery. There are issues related to the feasibility of randomization and blinding in surgery, ethical issues, standardization of the procedure, variations in surgical performance and variations among patients.

The method of assessing outcome in surgery needs to be tailored to each patient and generalization of the results is therefore difficult for surgery. Pattern Recognition (PR) provides tools for the assessment of surgical outcome for individual patients, and it allows for handling of outliers and individual patients and does not set the same restrictions on the data collection procedure as the RCT framework. PR could therefore provide a pragmatic next step towards data intensive operating room with evidence based support for surgeries.

2.1 Introduction

Evidence-based medicine (EBM) is a scientifically validated methodology developed to help clinicians make decisions based on scientifically valid evidence and results [Dex05]. Randomized Controlled trials (RCTs) [Ree07] is the most commonly used form of EBM. RCTs are a collective study design which allows researchers to scientifically measure safety and efficacy parameters of specific treatments (e.g. drug, diagnostic, device, and therapy protocols) within a group of patients. The *safety* of a treatment is determined by the observed adverse effects of a specific treatment during the trial. The *efficacy* is the ability of a treatment to reproduce a (hopefully desired) effect within the test subject group under controlled (RCT) conditions. The *effectiveness* is the ability of the treatment to produce the same effect in the real world; under less controlled or completely free conditions [Dex05; Mur04].

2.1.1 The RCT framework

In RCTs, a group of patients with specific symptoms, but otherwise healthy, are randomly split into two groups: an experimental group and a control group [Per08; Sol95]. The experimental group actually receives the treatment under investigation, whereas the control group receives the placebo (fake) treatment or the conventional treatment. Safety and efficacy data is gathered from patients within a predefined period. This data can include vital signs, concentration of the drug in the blood, and the improvement of the health of the patient. To ensure scientifically valid results, strict requirements on the trial design and data collection need to be fulfilled. As will be discussed throughout this paper, those requirements are not practical for all types of clinical treatments [Joh94], especially for invasive treatments.

2.1.2 Methodological challenges in surgery

It is hard for invasive treatments to conform to all requirements for RCTs, which include randomization, (double) blinding and placebo-control [Chu99; Mar03; Sch09; Far10]. Consequently, RCTs are very seldom applied to surgery [Kra05]. Fortunately there are outcome indicators for the performance of surgery available (e.g. injuries or infections after surgery), which can be used to evaluate the safety and efficiency of surgical procedures [Rev90]. If surgical research questions cannot be effectively addressed in a RCT study design, a study design should be used to allow for evidence-based decision support in surgery from the available perioperative data [Mar03; Bro08; Sch09].

To measure the safety and effectiveness of medical treatments, new individualized methods, while not yet widely used, are proposed to advance the field of clinical trial design [Far10; Dui07a; Tib96]. Some of the approaches build on prior knowledge, such as the Bayesian approach [Per08], rather than viewing each trial

in isolation as is the case in RCT studies. These methods also aim to avoid incorrect results, using advanced tools to compensate for biased data. Furthermore, they require fewer samples -and thus costs- and lead to rapid results.

2.1.3 Goal of the study

The objectives of this study are to examine the limitations of applying the RCT framework in surgery, and to investigate the feasibility of using Pattern Recognition (PR) as new approach for building evidence in surgery from -preferably available- perioperative data. After presenting the limitations of RCTs, this paper examines how surgeons can be supported in making decisions for individual patients based on historical data (i.e. prior knowledge). It is explained how data can be processed to support the surgical team in its decision making using PR techniques.

2.2 Applying the RCT framework for surgery

Although RCTs are a very powerful framework to improve the safety and effectiveness of clinical treatments, they have major limitations when applied in surgery [Chu99; Mar03; Sch09; Dui07a; Tib96; Far10]. This Section discusses the limitations of applying RCT in surgery for both the data collection process 2.2.1 and the statistical analysis 2.2.2. Finally, Section 2.2.3 discusses how these limitations are reflected in literature on surgical RCTs.

2.2.1 Data Collection: Limitations of satisfying RCT requirements in surgery

An invasive procedure (i.e. surgery) is not easily reproducible over a large population of patients. Unlike other treatments (e.g. drugs), invasive procedures can be (and usually are) adjusted to the specific situation of the patient. Even if the intention is to execute the procedure identically for all patients, there are sometimes external factors that introduce bias in the results of the study: the surgeon's expertise, the patient status, the anaesthetic, and the surgical team. Finally, it is expensive to repeat invasive procedures over a large population of patients making any study very expensive.

In order to apply RCT for surgery, one should either find a surgeon to perform all the procedures or assume that surgical skills do not significantly contribute to the surgical outcome. The latter is unfortunately not always a valid assumption [H.00]. Also, patients must be randomized and blinded to either the experimental or control treatment. This Section discusses the limitations of applying these constraints in surgery, assuming that within a surgical RCT a single surgeon (and in fact only one surgical team) performs all procedures:

- Randomization: It is much more difficult to randomly assign test subjects to 2 surgical interventions than 2 drugs, as both patients and surgeons most likely will have a preferred choice. Also surgical procedures are almost always permanent. This point may be of particular concern if a medical therapy is being compared to a surgical procedure or when two surgical procedures differ in magnitude or invasiveness (e.g. open vs. laparoscopic surgery). Here any null hypothesis (the null hypothesis is usually: no significant difference in clinical results) will be rejected on a priori grounds. In many cases, surgical RCTs are performed to compare conventional and new surgical techniques. However, a randomized trial may be impossible in many surgical scenarios for ethical reasons, because of the impact and risk of a surgical procedure for even healthy test subjects. It is most unlikely that any ethics committee would sanction the random allocation of test subjects to cardiac transplantation [Bla96].

- Blinding: Blinding is particularly difficult in surgery. The constraint of blinding is often overwhelmingly hard to overcome in surgical trials. It is much easier for patients to be oblivious to which medication was given to them than to not know which type of surgery has been performed on them [Nor03]. Double blinding is a larger challenge yet. Therefore one must assume that both patients and surgeons will know which surgery they are allocated to. The lack of blinding may be minimized by choosing a "hard" outcome indicator, such as mortality or morbidity [Sol95], which cannot be influenced by personal bias. On the other hand, if the outcome indicator is largely subjective (e.g. a change in symptoms or quality of life), lack of blinding will most likely bias the results. In this case blinding can sometimes be realized by an independent assessor who is unaware of the patient's treatment group.

- Placebo (sham) surgery: is a faked surgery performed in the control population to assess the effect of the intervention under study. Many ethicists reject sham-surgery [Dek01; Mac99], other maintain that such trials are ethically acceptable but should conform to certain restrictions that puts very strict requirements on the trial [Alb02; Mil03]. Such restrictions include that the research question cannot be answered by any other form of trial or study and that the risk of such procedure can be kept to a minimum.

- Equipoise: One of the major factors limiting surgical trials is the lack of community equipoise in surgery. Clinical equipoise is an ethical dilemma introduced by Freedman [Fre87], which can be paraphrased as: "genuine uncertainty within the expert medical community on the optimal approach for a certain medical condition". Equipoise allows clinical investigators to continue a trial until they have enough statistical evidence to convince other experts of the validity of their results, without a loss of ethical integrity on the part of the investigators. If more than 70% of the medical experts

favored one of the treatment options, we considered that community equipoise was lacking 70% [Sol95]. Johnson et al. [GHO08] proposes to apply equipoise in a retrospective way to patients where there was disagreement over their treatment.

- Timing: The issue of timing of trials is difficult. Most surgeons agree that new surgical techniques would change significantly within first few years [Sol95]. In any surgical procedure there exist a learning curve and modifications to the new surgical technique are made frequently. By including these early patients, one would almost certainly bias the results against the new procedure. The introduction of laparoscopic cholecystectomy and the initially high rate of common bile duct injuries is a good example of this [Mur04]. On the other hand, it may be difficult and quite unnecessary to initiate a trial when the procedure is widely accepted by both the patient and the surgical community.

- Multi-centre: Frequently, surgical trials rely on results from single institutions or from multiple centres within one health area or organization. However, there may be inherent selection biases in referral patterns to the institution ('centre effect') and study participation; as well as limitations posed by sample size and completeness of data [Hal05]. Please note that this constraint is contradictory to the earlier assumption of using an identical surgical team for all procedures.

2.2.2 Statistical analysis: Limitations of applying RCT hypothesis testing in surgery

After a RCT is conducted, the data is statistically analysed, mostly by means of hypothesis testing [Per08]. For each collected parameter, a research question is formulated in the form of two hypotheses which are mutually exclusive, i.e. if one is true, the other must be false. These are often characterised as the null hypothesis H_0 and the alternate hypothesis H_a. Typically H_0 is chosen to reflect the situation that there is no statistically significant difference in a single parameter (i.e. representing safety or efficiency) between the control and experimental groups [Wal11]. For example, the null hypothesis could state that there is no difference in death rate between patients who took aspirin daily and those who did not, the alternate hypothesis would automatically state that there is a difference [Dex05; Per08].

To perform hypothesis testing, the probability density function (pdf) of the control group has to be approximated. Often the pdf is approximated by a standard normal distribution with $\mu = 0$. This approximation is justified using the Central Limit Theorem (CLT) [Eck10], which states that when you have many, small, independent, random variables, then their sum is distributed as a bell-curve (i.e. normal distribution). In order to use the CLT, the data must be collected

within the requirements of Section 2.2.1. Assuming the data is collected in line with the requirements of Section 2.2.1, there are still limitations of applying the RCT hypothesis testing in surgery:

- Outliers in surgery: An outlier is an observation that does not follow the pattern of the majority of the data [Tax98]. Fitting a normal distribution on the data emphasizes the effect of average patients and understates outliers [Bla96]. This is of no concern if one is interested in the ordinary behaviour of a therapy. However, in surgery the information about outliers (specific individual patients) is also very relevant. Outliers such as complications and adverse events define directly the negative outcome of surgical interventions. Those outliers are often clinically important and may correspond to surgical errors; hence they are worthwhile to be flagged and analysed, especially in surgery where complications are very common and hard to predict [DB07; Gri08].

- Factor-dependency in surgery: Many measurements seem to fit a normal distribution, especially when enough data is used as the case is in the RCTs of most medical treatments. Because this assumption tends to work well most of the time, it is usually taken for granted in many domains. A major reason why CLT fails is that the individual factors of a given study are not independent and therfore correlated [Tal11; Bla96]. In this case the data will not fit a normal distribution well. Surgery is one of those domains where variables may not be independent and other factors may influence the outcome (e.g. surgical skills) which result in considerable bias when evaluated in such a strict hypothesis testing framework [Dui07a; Tib96; Far10]. To avoid bias in surgical RCTs, a very high sample size is needed, resulting in much larger scale RCTs than the case in other disciplines of medicine.

2.2.3 Surgical RCTs in literature

Even if the above challenges could be overcome, 60% of the surgical questions cannot be addressed in an RCT trial [Noc10]. Most of "surgical" RCTs evaluated drugs in a surgical setting, less than 25% of the published trials involve a surgical procedure in the trial [Chu99]. Nevertheless, in some surgical fields, surgeons conducted a number of surgical RCTs. For example: in the published RCTs related to digestive surgery, 84% evaluated the core surgical procedure by comparing surgery versus drug, two different surgical strategies or a changed important step of the procedure (e.g. anastomosis, drainage) [Chu99]. Problems with RCTs in surgery are related to very different patients, feasibility of randomization and blinding, the learning curve, standardization of the procedure and patients' and surgeons' equipoise [McL99]. Furthermore, as surgical procedures are often much more expensive and risky to perform, especially in comparison to drug treatments, large scale RCT studies are much harder to perform.

The number of RCTs in surgery is, given the above constraints, understandably rather limited. Only 3.4% to 7% of the publications in leading surgical journals are randomized trials [Tib96]. Of these trials, only 20% to 50% deals with the comparison of 2 or more interventions. Despite the fact that clinical surgical trials are mainly based on nonrandomized studies, however, there is no agreement on good alternatives to RCTs in surgery [Chu99].

2.3 Pattern Recognition (PR) as a new perspective for EBS

In order to use EBM in surgery, new techniques need to be found to deal with the individual nature of surgeries. This section proposes PR to measure safety and effectiveness parameters in surgery. The need for randomization and blinding hold for any statistical approach including PR. However, PR provides an alternative for building classifiers, based on historical samples from other patients or from other clinical trials, to support surgeons in decision making for individual patients.

2.3.1 Introduction to PR for surgery

Pattern Recognition (PR) enables, through classification of pre-, intra- and post-operative data, the prediction of surgical outcome. This prediction (i.e. classification) is based on historical patient data. This predicted outcome can support the surgeon in making evidence based decisions for new patients. The data is pre-processed into specific data-points, which are known as feature vectors and serve as input for the classification algorithm (i.e. classifier), that assigns the sample to one of the given classes [Tes99; Fuk90]. The classes are all possible surgical outcomes (e.g. occurrence of complications: injured, inflamed, infected, none). In order for PR to classify a patient, the patient must first be described in a way that the PR algorithm can understand [Tes99]. Given a new patient to analyse, a PR system must first generate a description of it in terms of a vector of P features (i.e. the pattern) $\mathbf{x} = (x_1, ..., x_P)$. When the patient is known to belong to one of C classes, an additional class label $y \in \{1, ..., C\}$ is defined.

PR aims at constructing a classifier $f(\mathbf{x})$ that can predict the class based on the input features. In order to construct a classifier, a labelled training set $X = \{(\mathbf{x}_i, y_i), i = 1, ..., N\}$ is needed to train the classifier. With this training set, the decision rule(s) within the classifier are defined such that the probability of misclassifying any $\mathbf{x}$ is minimized. After evaluating the classifier, it can be used in a clinical setting to classify features representing new patients and predict the most likely class for this patient.

To illustrate the PR framework, consider laparoscopic cholecystectomy as the intervention. Patient demographic data can be used as features to predict occurrence of major bile duct injury (i.e. the occurrence of the injury is the first class, the second class is its absence). From available patient data, a classifier can be trained to classify laparoscopic cholecystectomy patients into class $y = 1$ (major

bile duct injury) or class $y = 2$ (no major bile duct injury). Once the classifier is trained and shows a good performance, it can be used in a pre-operative setting to predict the most likely outcome for new patients. Predicting the occurrence of complications before surgery can aid surgeons do decide whether to proceed with a minimally invasive approach, to perform an open procedure or to make a referral to a more experienced surgeon [Bou11c].

2.3.2 Branches of PR in surgery

PR in surgery (and equally for other application domains) can be deployed in three levels.

- Basic PR: A classifier that classifies instances into a pre-defined class
- Branch 1:Basic PR plus the detection of outliers that do not belong to a pre-defined class
- Branch 2: Basic PR plus the ability to cope with features evolving over time (i.e. non-stationary features)

Basic PR: Classification into a pre-defined class

A classifier is constructed from the training dataset X. This classifier is used to classify new feature vectors which represent new patient cases into one of the pre-defined classes. At this level, it is assumed that all instances belong to one of the initially defined classes. Further, it is assumed that the feature set is derived from stationary data sources (i.e. the features do not evolve over time). An example is to predict wound infection by patients after ventral hernia repair using features such as demographics, perioperative risk factors and operative characteristics. For any classification task we need data from two or more classes, in this case a group of patients with no wound infection after surgery (class 1) and the group of patients with wound infection after surgery (class 2).

Once the classifier is trained with the available labelled data, the most likely class for new patient cases can be predicted. The decision boundary between the two classes is formed by the feature vectors that are equally likely to belong to either class. New patient cases are classified according to which side of this decision boundary they are located, as is illustrated in figure 2.1. Different types of classifiers can be used, ranging from simple classifiers such as Linear Discriminant Classifier (LDA) and nearest-neighbour classifiers (NN), to more complex classifiers such as Support Vector Machines (SVM).

Under-fitting and over-fitting are important issues for any classifier. Figure 2.2 illustrates the problems of over-fitting and under-fitting. Assuming sufficient samples of training instances are available to train the classifier from both groups, the challenge is to construct a classifier that fits the data correctly. Using a simple, non-flexible classifier for complex data can lead to under-fitting, as illustrated in

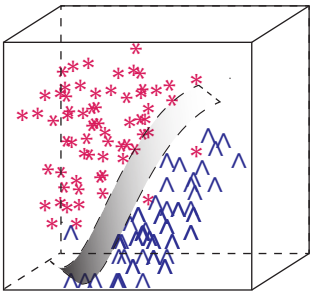


Figure 2.1: *Classification by means of separating surfaces between two classes*

Figure 2.2(a). The classifier is incapable of following the intricate optimal decision boundary, resulting in suboptimal classification performance. On the other hand, using a too complex, too flexible classifier with small amounts of data, may lead to over-fitting. Here the classifier adapts to the noise and to the structure in the data that may not represent the class. This results in a poor classification performance for new patient cases, as illustrated in Figure 2.2(b). The trade-off is obviously somewhere in between, a classifier that adjusts just enough to the data, without over-fitting or under-fitting to the details of the data as illustrated in Figure 2.2(c).

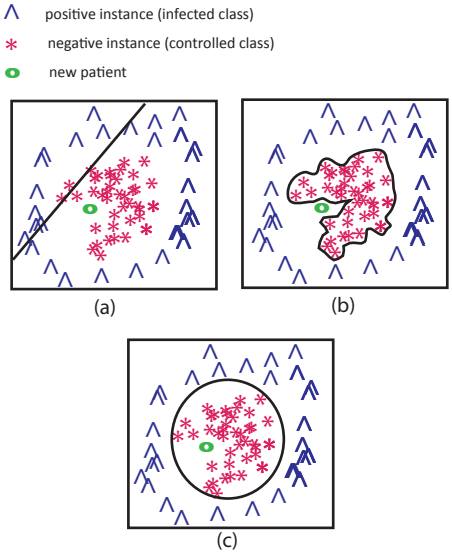


Figure 2.2: *Under-fitting and Over-fitting*

Although the classifier can show very good performance for all patients in the training dataset, this is no guarantee that the classifier will perform well for new datapoints. Thus the performance of the classifiers needs to be evaluated with new datapoints. Appendix A explains how to evaluate the performance of the trained classifiers using the available data.

Branch 1: Detecting outliers

The assumption that all data (training as well as new data) belong to one of the pre-defined classes, is often not true. When patients have developed atypical features for their class, or have been selected erroneously, the classification performance will be adversely affected. Note that this assumption is always made in RCTs. In figure 2.3 an example of an outlier is shown. A classifier is constructed that can correctly classify most of the instances correctly. However, the outlier in the lower right corner will be incorrectly classified as a positive instance (e.g. infected patient). In such a case we want the classifier to reject this object and label it as an outlier [Bis95; Hod04a].

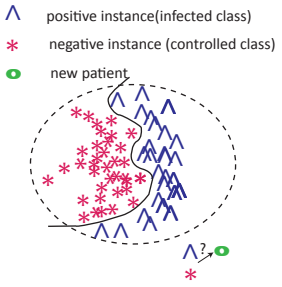


Figure 2.3: *Example of an outlier*

In PR, there are different approaches that can be used to detect outliers with respect to one of the predefined classes [Hod04b]. One approach is to approximate a single (multi-dimensional) probability density function (pdf) per class, using for instance a normal distribution. Applying an arbitrary cut-off threshold, outliers can be identified. A more sophisticated approach is to represent the class pdf as a (weighted) sum of densities. A sum of multiple normal density functions can be represented in a Gaussian Mixture Model which is used to represent density functions where the CLT does not apply, and that need more complex density functions [Bis94]. As with the single pdf approach an outlier is detected when it has a lower probability of occurring than a predetermined threshold. Other approaches measure the distance between features within a class to detect outliers. Tax et al. [Tax98] compared the distance between nearest neighbours within the

same class to detect outliers. Those approximation-free approaches are more flexible and reliable to apply with less heterogeneous data such as surgical data.

Branch 2: Classifying time series

Teh basic PR assumes a stationary feature set: the data is independent of time and can thus be processed disregarding any temporal relationships. However, in many cases the feature set is a non-stationary process derived from data evolving over time. Many patient characteristics are time series of measurements such as blood pressure, heart rate, cardiac rhythm, expired CO2 and laparoscopic video. Time series are commonly used to classify critical events during interventions. Note that, in case of time series we deviate from the classical PR framework. There are not two or more distinct classes for the entire procedure, instead there are many time events taking place that need to be classified. For example, using laparoscopic video, we want to detect the common bile duct transection (cutting across the wrong duct) during laparoscopic cholecystectomy. To detect the common bile duct transection, first we need to detect the clipping step during the intervention. For this classification task we need to use non-stationary data (video frames) to detect the clipping step.

Techniques for interpreting non-stationary sources of data are not as developed, nor as established, as those for a classical static problem [Bis95]. There are, however, two approaches to deal with the stationary nature of the feature set in a classification problem.

The first approach divides the time series into separate, smaller series of successive features, to enable the use of analysis tools developed for stationary signals such as classification [The09]. Each window consists of a finite number N of features (e.g. 25 frames = 1 second of video). During this short time interval, the signal is assumed to be stationary. Choosing the right length of windows is a crucial step and problem-dependent. The window must be long enough to capture the necessary information for the classification task and short enough to guarantee the (approximate) stationary of the signal. After choosing the window, the classical classification in the approaches described above can be applied for each window. Consider that we want to detect the clipping step during a laparoscopic cholecystectomy using laparoscopic video by tracking surgical tools, this approach classifies each window separately whether it is a clipping step or not. Note that in the clipping step different tools are used: clip tang, dissecting device and scissor. Those tools can also be used in other steps of surgery. The fact that at the window level two different surgical steps are composed out of the same elements (i.e. surgical tools) can make them indistinguishable when using this approach.

The second approach makes use of Hidden Markov Models (HMM) which encompasses the idea of the first approach and uses not only the class-specific characteristics (e.g. surgical tools), but also their relative order (e.g. first dissection, then clip placement, then cutting). This model allows representing time series as a series of transitions from one state to another (e.g. surgical steps)

by means of a transition matrix, and in the meantime also learn the state model that best represents each discovered state [Bou11b]. Using this approach the surgeon is provided with more detailed information on the likelihood of occurrence of complications. By monitoring the data during surgery the system might even be able to warn in case of imminent complications or significant deterioration of the condition of the patient.

2.4 Perspectives of applying PR in surgery

To avoid medical errors in surgery, PR can be used to support surgeons in making evidence-based decision tailored to individual patients. PR can provide tools to predict outcome indicators from historical patient data about the safety and effectiveness of the surgery. This Section discusses the prospects and challenges for using PR in surgery.

2.4.1 Prospects of using peri-operative data in surgery

Today's operating rooms (OR), post anaesthesia care units (PACU) and intensive care units (ICU) generate vast quantities of data. Sophisticated monitoring equipment performs continuous measurements of multiple physiological parameters, on a high frequency, that require timely and context sensitive analysis in order to sustain effective decision support [Sam06]. There is lot of unexploited data about surgical interventions that can be used (e.g. for adverse event prediction). This data is not collected in line with RCT requirements and is currently only very seldom exploited for evidence building.

Before the start of the surgery, patient data is already being collected. It can include demographic information, clinical history of the patient, current and past medication, measured vital signs, laboratory data, radiology examinations and nursing records. During surgery it has become routine to measure the patient's blood pressure, heart rate, cardiac rhythm, expired CO_2 and temperature. In laparoscopic surgery, the laparoscopic video can be used to track surgical tasks during the intervention. After the surgery patients are transferred to the PACU or the ICU units, O_2 information and much other information can be recorded. Because so much data is available, it is important to identify those variables most important to predicting adverse (i.e. safety) and favourable (i.e. effectiveness) outcomes. Searching for meaningful predictors of adverse and favourable outcomes is an important challenge to improve surgical treatment for patients.

pre-operative data: During pre-operative planning demographics and comorbidities are collected daily from patient's pre-operative history and physical examinations. Pre-operative data can also include recent laboratory values. Moreover, pre-operative imaging is occasionally used to plan the surgical strategy for localization and size assessment purposes. This data is usually not used for EBM because it does not satisfy the RCT requirements. There is a large op-

portunity to use this readily available pre-operative data to predict intra- and post-operative outcomes using PR.

intra-operative data: The growing availability of measurement devices in the operating room (OR) enables the collection of intra-operative data on the surgical workflow and the condition of the patient during surgery. Many hours of endoscopic video are produced that can be used to log surgical events. Moreover, physiological data about the patient's blood pressure, heart rate, cardiac rhythm, expired CO2 and temperature can be recorded. Also anaesthesia records are kept. Furthermore for robotic surgery data is recorded about the motion of instruments and their interaction with human tissue. Most of this data is not used, also because it does not satisfy RCT requirements. Intra-operative data is best suited for time series analysis to predict significant events during surgery.

post-operative data: Directly after surgery, patients are usually transferred to the post anaesthesia care unit (PACU), which is specifically designed to provide care for patients recovering from anaesthesia. In the PACU, patients are monitored continuously so that any difficulties that develop as they emerge from anaesthesia are quickly recognized. Many vital signs are monitored, like blood oxygen saturation, level of consciousness, independence of breathing and ability to make voluntary movements.

When patients are deemed too unstable for the PACU, they are transferred to the intensive/invasive care unit (ICU) instead. The ICU is, like the OR, a very data-rich environment. Monitors, as well as therapeutic devices (such as mechanical ventilators, syringe- and infusion pumps for drug and fluid administration, or renal replacement therapy machines), generate data on a continuous basis. Blood samples for laboratory analysis are drawn several times a day, and microbiology sampling occurs several times a week. Doctors and nurses write progress notes several times a day. Drug prescription and delivery is changed and charted more than daily.

2.4.2 Measuring safety from peri-operative data

Safety outcome indicators

Surgical safety depends heavily on the surgical speciality. What is considered a safe surgery is highly dependent on the characteristics of the intervention: the type of surgery, the seriousness of the condition it is aiming to treat, the experience of the surgeon in performing this type of surgery and the condition of the patient.

Most safety indicators are related to intra-operative complications (i.e. adverse events during surgery). Table 2.1 describes possible intraoperative complications after different laparoscopic interventions as described by Perugubu et al [Per01]. The most serious intra-operative complication for laparoscopic cholecystectomy is biliary injury. For laparoscopic hernia repair, bladder injury is an example of intra-operative complications. For patients undergoing laparoscopic colectomy, intra-operative complications include enterotomy, mesenteric bleeding

and ureteric injury.

Table 2.1: *Example of intraoperative complications from different laparoscopic procedures*

Laparoscopic speciality	intra-operative complications
Cholecystectomy	major bile duct injury
Antireflux surgery	perforation of either esophagus or stomach splenectomy pneumothorax
Inguinal hernia repair	enterotomy mesenteric bleeding ureteric injury

Measuring safety

Measuring safety is a challenging task in surgery. One of the main barriers in measuring safety in surgery is the lack of standardization. Standardization is difficult in the domain of surgery because surgeons may vary in their experience with and ability to perform a surgical technique, there can be individual preferences in performing the procedure, and technical modifications may occur as the procedure evolves. Moreover, differences in perioperative and postoperative care may also impact the outcome [Wol07].

The first issue (who performs the procedure) is analogous to assessing compliance in a medical trial. It would be appropriate not to limit surgical participation to expert surgeons from the same school. When surgeons from different schools perform the same procedure they need at least to agree on the how to perform the procedure (i.e. consensus). It may not be necessary that there is agreement concerning all the technical aspects, but there should be consensus on those details deemed to be important [Nor03]. Hence, evidence-based studies for surgery need to allow the detection of the surgical consensus regardless of how it is executed by different surgeons.

Consensus in surgery exists generally only on a high level, and it often disappears when one closer examines the procedure. Due to the high amount of uncertainty in surgical decision making, surgical tasks show large amount of variations on a low level in performing a procedure. Therefore, measuring only the consensus of surgeries cannot help in solving difficult and rare surgical situations (i.e. outliers) [Dui07a]. Those outlier situations are very important but do not produce sufficient samples for any formal EBM study. Therefore, we need to detect and understand the nature of outliers during surgical interventions.

Finally, any surgeon has his own learning curve in mastering novel surgical techniques (e.g. new instruments or a robot). As the surgical technique is novel, modifications are made to it more frequently than conventional techniques. It is important to continuously measure this learning curve by following early patients.

We summarize the challenges for measuring surgical safety as follow:

- Challenge 1: Predict intra-operative complications from readily available pre-operative patient data.
- Challenge 2: Measuring the execution of surgical consensus (or critical steps) during interventions regardless of the level of variance of its execution by different surgeons and for different patients
- Challenge 3: Measuring bias in executing the same surgical consensus (or critical steps) by different surgeons
- Challenge 4: Detecting outliers (e.g. adverse events) during surgical interventions as a first step towards formal outlier management
- Challenge 5: Monitoring the learning curve and the effectiveness of new surgical procedures

2.4.3 Measuring effectiveness using peri-operative data

Effectiveness outcome indicators

Measuring effectiveness is particularly hard in surgery. For many surgeries the outcome becomes apparent only years afterwards, for others the results may be instantaneous. In general "hard" outcome indicators such as mortality or morbidity, a change of symptoms or quality of life will take years to be noticed. Such outcome indicators can only be collected in long-term systematic initiatives and focus mostly on high impact surgeries, such as heart bypass surgeries [UK12b].

Immediate post-operative outcomes, is an effectiveness indicator which can be measured directly after surgery. It may either be general or specific to the type of surgery undertaken, and should be managed considering the patient's history and the surgeon's surgical expertise. Most of the post-operative complications become apparent between one and three days after surgery [UK12a] and are related to both safety and effectiveness of the surgical intervention. General post-operative complications include fever, wound infection, embolism, haemorrhage, respiratory complications, urinary problems, increase in blood pressure, blood loss, nausea and vomiting. Infections are one of the main causes of morbidity in abdominal surgery and can appear within the first weeks after surgery. Specific post-operative complications are related to specific surgical procedures and can include diathermy burns and thrombo-embolism.

Measuring effectiveness

With a large amount of new instruments and devices introduced to surgery, surgeons frequently decide whether to adapt their surgical techniques to this new technology. Therefore, it is crucial to compare the effectiveness (and also the safety) of the new surgical technique with the conventional one. In order to have sufficient data available to make this decision, it is important to follow patients and log their post-operative outcomes, surgical consensus and any outliers.

Nevertheless, it is important to know what happened in the intra-operative stage to retrieve the cause of the post-operative complication. Accordingly, it is important to measure the surgical consensus as well as the outliers during surgical interventions as described in Challenges 1 and 3. Hence, measuring effectives shares the same challenges to measuring safety with the ability to:

- Challenge 6: Compare the effectiveness of new surgical techniques when compared to conventional techniques.
- Challenge 7: Predict post-operative complications from readily available (pre- and intra-operative) patient data.

2.4.4 Measuring efficiency using peri-operative data

For hospitals, excessive patient waiting times can create systemic bottlenecks and ultimately menace patient safety. The most effective way to bring more efficient care is the employment of EBM whenever possible [Ost10]. The goal of surgical efficiency research is to minimize required resources and related costs, while maintaining patient safety.

Various performance criteria are used to evaluate surgical efficiency. Cardoen et al. [Car10b] distinguished different performance measures. Among the most heard issues in surgery is the long waiting lists, which justifies many studies aiming at decreasing waiting times between surgeries (i.e. OR idling time), surgeon's waiting time and operating room overtime. Hence, increasing the throughput of the hospital by increasing the number of treated patients.

Another important parameter of effectiveness is the utilization of expensive units such as the OR. The utilization rate of the OR has been the subject of different studies. Both (under-) and (over-)utilization of the operating room represent unnecessary and unstable costs for hospitals [Car10b]. Besides the operating room itself, the occupancy of closely connect resources to the OR should be considered, namely the PACU and the ICU. OR capacity problems can also be caused by unexpected events in the PACU, the holding area or the ICU. When throughput of the operating room is improved, closely connected resources must keep up, in order to achieve an overall capacity benefit.

In most hospitals, patients move through their operative day in a predetermined, linear way; they start at registration and finish in the recovery room. To allow a dynamic scheduling of surgeries different statistical approaches can be

used: offline approaches (i.e. before schedule execution) and online (i.e. during schedule execution) [Car10a]. Using PR techniques dynamic schedules can be developed that lead to smooth surgical resource utilization without peaks by reducing intra-operative, flow-time, wait-time and operative time.

We summarize the challenges for measuring surgical efficiency as follow:

- Challenge 8: Efficient planning of surgical resources using pre-, intra-, and post-operative data.
- Challenge 9: Offline prediction of surgical resource occupancies: surgical time, recovery time, possible complications.
- Challenge 10: Online prediction of surgical resource occupancies: surgical time, recovery time, possible complications.

2.5 Conclusion & Discussion

2.5.1 Discussion

Surgery is a skill-dependent, multistep procedure. This makes evidence based studies in a traditional RCT framework difficult to be designed. However, evidence-based justification of surgical practice is becoming increasingly relevant to avoid adverse events. This paper proposed pattern recognition as an alternative approach which allows us to estimate the safety, the effectiveness and the efficiency of a surgical treatment for individual patients, using the available biased, noisy and incomplete data. Although it does not provide the same level of evidence as an RCT, it allows for variations in surgical practice and patient anatomy. This section discusses the characteristics of surgery, the qualifications of using PR for evidence based surgery and the alternative of using observational studies for EBM.

On the characteristics of surgery

Surgical interventions are tailored to individual patients and surgical teams. RCTs are, however, generalized models aimed to capture the average effect of the treatment under study. Accordingly RCTs understate individual variations caused by, for example, the patient's history. This average effect is achieved by averaging the effect of the same drug on a sufficiently large patient population.

The variations in surgery bring about another issue, which is the gap between efficacy and efficiency. Although this gap can already be quite large for drugs and other clinical treatments, it is especially prominent for surgery. This is caused by need for reproducible treatments versus the real-world practice of surgery with a wild variety in surgical team experience and expertise, different instruments, anaesthetics, patient anatomy, etc.

On the qualifications of applying PR vs. RCT to provide evidence based surgery

PR allows the prediction of surgical outcomes (especially when related to safety and effectiveness) by classifying patients based on their individual characteristics (i.e. features). It builds on prior knowledge, rather than viewing each trial in isolation as is the case in a RCT. In PR, the surgical outcome forms the classes of the classification problem. For example, when the surgical outcome under study is surgical injury, we have a two-class problem: the null class contains the samples of patients without surgical injury, and the alternate class contains the samples of patients with surgical injury. The patient's characteristics make up the feature set which serves as the input to a classifier (i.e. set of decision rules) which assigns the sample to one of the given classes.

PR also allows to perform hypothesis testing, without assumptions for the distribution of the null class. Therefore, it is less tied to the requirements of Section 2.2.1. PR can always be used, however the heterogeneity of the data is a relevant factor in the overall result. These requirements are also important for PR, they are, however, not required to start the analysis as is the case in RCT.

PR allows detection of outliers while RCT focuses on generalized cases. PR makes no assumption about the distribution of the classes, instead any distribution of the classes can be used which allows outlier detection using approaches discussed in Section 2.3.2

Finally, PR has no hard restrictions on the size of the data needed for building a classifier, the analysis can be started with any reasonable sample size. Moreover, through learning curve analysis one can assess if more samples will improve the performance.

On the alternative: observational studies

Observational studies is another form of EBM study designs where isolated trial studies are performed without using prior knowledge. It include cohort studies, case-control studies and cross-sectional studies. In an observational study the researcher only observes without performing any clinical intervention. This is typically used when RCTs are deemed unethical, or if the treatment to be studied does not fit the strict requirements of an RCT. Cohort studies are used to study incidents, causes, and prognosis. Because cohort studies consider events in chronological order, they can be used to analyse cause and effect. Cross sectional studies are used to determine prevalence. They are relatively quick and easy to perform but lack distinction between cause and effect. Case controlled studies compare groups retrospectively. They seek to identify possible predictors of outcome and are useful for studying rare diseases or outcomes. Observational studies usually provide (much) less compelling evidence than an RCT study.

Although observational studies are a realistic choice, particularly when an RCT would be impractical, only the most basic statistical analysis tools are used (e.g. ratio of probabilities: odd factor, relative risk) to draw conclusions from

the observed data. In case of cohort studies, the relative risk is calculated as the ratio of the probability of the event occurring in the experimental group, and the control group. Considering the high amount of bias in observational data, due to lack of randomization and other RCT requirements, the basic statistical analyses are not sufficiently powerful to extract the best results from the observational data. [Kno08; Gro07; Dex05; Mur04; Mar03; Bro08; Tib96]. Instead, PR provide more advanced tools to extract prior knowledge from the complexe surgical data and use it for future predictions. PR can also be used as a complement to standard statistics in observational studies to cope with the limitations of its data, where the number of variables, the size (number of data points), and the quality of the data (missing data, inaccurate transcriptions) would make standard statistical methods ineffective.

2.5.2 Conclusion

This paper discusses limitations of RCTs and the perspectives of using PR for individualized evidence based surgery. There are factors that contribute to the outcome in surgery which result in considerable bias when evaluated in a RCT framework. Also other, less relevant issues are prevalent: ethical issues, lack of standardization of the procedure, variations in surgical performance and variations among patients. This in combination with high costs and inherent risks of surgery lead us to the conclusion that RCTs are not practical for the vast majority of surgical procedures.

Fortunately Pattern Recognition (PR) techniques provide alternative tools for evidence based surgery. PR can be used for small datasets, allows for handling of outliers and individual patient cases and does not set the same restrictions on the data collection procedure as the RCT framework. Unfortunately PR does not provide the same quality of evidence as an RCT study, without setting the same requirements as an RCT study. However less evidence is for many cases quite acceptable given the individual and informal tradition of surgery. PR can provide a pragmatic next step towards data intensive OR with evidence based support for surgeries.

Chapter 3

Preoperative: prediction of intra-operative complexity

L.Bouarfa, A. Schneider, H. Feussner, N. Navab, P.P.Jonker, J.Dankelman
Published in the Journal of Artificial Intelligence in Medicine, (2011)
under the title "*Prediction of intra-operative complexity from preoperative patient data for laparoscopic cholecystectomy*"

abstract

Objective: Different reasons may cause difficult intra-operative surgical situations. This study aims to predict intra-operative complexity by classifying and evaluating preoperative patient data. The basic prediction problem addressed in this paper involves the classification of preoperative data into two classes: *easy* (Class 0) and *complex*(Class 1) surgeries.

Methods and material: preoperative patient data were collected from 337 patients admitted to the Klinkum Rechts de Isar hospital in Munich, Germany for laparoscopic cholecystectomy (LAPCHOL) in the period of 2005-2008. The data include the patient's body mass index (BMI), sex, inflammation, wall thickening, age and history of previous surgery, as well as the name and level of experience of the operating surgeon. The operating surgeon was asked to label the intra-operative complexity after the surgery: '0' if the surgery was easy and '1' if it was complex.

For the classification task a set of classifiers was evaluated, including linear discriminant classifier (LDC), quadratic discriminant classifier (QDC), Parzen and support vector machine (SVM). Moreover, Feature-selection was applied to derive the optimal preoperative patient parameters for predicting intra-operative complexity.

Results: Classification results indicate a preference for the LDC in terms of classification error, although the SVM classifier is preferred in terms of results concerning the area under the curve. The trained LDC or SVM classifier can therefore be used in preoperative settings to predict complexity from preoperative patient data with classification error rates below 17%. Moreover, feature-selection results identify bias in the process of labelling surgical complexity, although this bias is irrelevant for patients with inflammation, wall thickening, male sex and high BMI. These patients tend to be at high risk for complex LAPCHOL surgeries, regardless of labelling bias.

Conclusions: intra-operative complexity can be predicted before surgery according to preoperative data with accuracy up to 83% using an LDC or SVM classifier. The set of features that are relevant for predicting complexity includes inflammation, wall thickening, sex and BMI score.

3.1 Introduction

Pattern recognition is gaining increasing attention in the medical domain, as it has proven more effective than common clinical statistical tools in the prediction of clinical outcomes. Pattern recognition has been used for decades for segmentation purposes in medical imaging applications [Mai98]. Classification methods have recently been used for decision support in computer-aided diagnosis. For example, Lin et al. [Lin09] designed a diagnosis model for the treatment of liver disease using classification and regression trees, and Lee et al. [Lee10] developed a computer-aided diagnosis system for evaluating pulmonary nodules using feature selection and a linear discriminant classifier (LDC).

The vast amount of preoperative patient data generated before surgery motivates the construction of pattern recognition tools that are able to improve the accuracy of predictions regarding complexity factors that may occur during surgery. Nevertheless, the use of surgical preoperative data for prediction has been overlooked in the literature. This challenge has recently been addressed by conventional clinical statistical approaches, which lack the power of pattern recognition when using ranking and classification algorithms as prediction tools.

3.1.1 Why predict surgical complexity?

Surgical complications have been associated with increased inpatient hospital costs [Dim04]. As a result, reducing complications has become a desirable objective for quality-improvement initiatives aimed at improving efficiency and safety in health care. Davenport et al. [Dav05] have shown that preoperative risk factors and surgical procedure complexity are more effective predictors of hospital costs than complications are. This dependence between risk and procedure complexity is to be expected, as these measures were designed to predict complications. Nevertheless, the use of raw preoperative patient data in predicting complexity factors has been overlooked in the clinical literature. In most clinical studies [Din04], complexity factors often are based on nothing more than surgical expertise. Complexity factors are pre-designed and classified according to the surgeon's knowledge about possible complications of a specific procedure. Moreover, in the clinical literature, surgical complexity is estimated in per procedure and not with regard to the relative ease or difficulty of a procedure for a given patient [Dav05]. Nonetheless, the literature does contain a small number of studies that assess the surgical complexity of individual procedures according to readily available patient data. Jenkins et al. [Jen10] used patient demographics to predict the operative time for ventral hernia repair. In most cases, however, these approaches are statistical, concentrating only on evaluating the significance of individual parameters as independent variables. They lack the power of pattern recognition tools in performing classification and selecting a subset of parameters for which the classification performance improves the most.

3.1.2 Complexity prediction for laparoscopic cholecystectomy: why is it important?

Laparoscopic cholecystectomy (LAPCHOL) is one of the most commonly performed surgical procedures worldwide [Sod10]. LAPCHOL is accepted as the gold standard in the treatment of symptomatic gallstones. Up to 700 000 LAPCHOL procedures are performed in the US each year [Sod10]. The preoperative assessment of complexity factors is needed for frequent procedures such as LAPCHOL in order to avoid complications and delays and to guarantee an efficient course of surgery.

This study aims to estimate the intra-operative complexity of LAPCHOL according to readily available preoperative patient data. Resource planning is a crucial topic in research on surgical efficiency. Tools are prepared in essentially the same way for all surgeries. In many cases, however, surgeons need advanced tools in case of complications. These tools can be prepared preoperatively for surgeries that are predicted to be complex. Furthermore, any complex procedure always involves the risk of conversion to an open procedure. This measure can be taken preoperatively to avoid intra-operative delays during surgery. The members of the surgical team can be considered another example of resources. If a surgery is identified as complex, it can be assigned to a more experienced team (including the surgeon, the surgeon-assistant or the operative-nurse), thereby allowing for a safer and efficient surgical procedure. Moreover, although the average time required for LAPCHOL is about 60 minutes, in practice, it may vary from 20 minutes to 5 hours, depending on the level of complexity [Dex06]. The variation in operative time is due to uncertainties regarding the complexity of the LAPCHOL procedure across a diverse patient population [Jen10]. Estimating the level of complexity beforehand may improve the flexibility and accuracy of preoperative planning. The scheduled time can be increased for complex procedures and decreased for easy ones, thus allowing flexible preoperative planning.

In complex surgical situations, the surgeon gets into dilemma weather to continue the intended operation, or to deviate from the planned procedure. Complexity estimation before surgery can aid surgeons in decisions regarding whether to proceed with a minimally invasive approach, perform an open procedure or make a referral to a more experienced surgeon. It may also be useful as an informative tool for communicating about details of the intervention with the patient and for explaining the various risks of laparoscopic and open procedures.

3.1.3 Goal and contributions

This study aims to delineate the relevant demographic and sonographic patient data that are predictive of the intra-operative complexity of LAPCHOL interventions.

This work contributes in several ways. First, it introduces pattern recognition tools for processing surgical data; in the medical literature, these data are usually

processed using common clinical statistical tools. Second, this study provides an evaluation of a number of classifiers based on a 337-patient dataset that was collected in the period of 2005-2008. Third, it provides an analysis indicating which set of preoperative features is effective for predicting intra-operative complexity, thus filtering out insignificant features. The study also measures objectivity bias in the assessment of surgical complexity by surgeons with various levels of experience. Finally, the study evaluates the conformity of the results with clinical practice.

3.2 Materials and methods

The preoperative features used in the experiments were collected in the period of 2005-2008 from $N = 337$ patients who had been admitted for elective LAPCHOL procedures, which involve removing the patient's gallbladder in case of symptomatic gallstones.

3.2.1 Dataset

All patients had received one or more preoperative sonographic examinations of the gallbladder. Table 3.1 presents the collected data, including patient demographics and sonographic features. Each feature is listed by name and data value, as encoded in the dataset. Demographic features include sex, body mass index (BMI) and previous surgeries. Sonographic features include wall thickening, size calculi and the presence of inflammation. Data were collected from 23 different surgeons, including their ID and level of experience as surgeons. To measure the actual clinical situation, the operating surgeon was asked to provide a score of intra-operative complexity following each surgical procedure: '0' if the procedure was easy and '1' if it was complex.

3.2.2 Method

Two steps are considered for classifying preoperative data. In each of these steps, the data are transformed through specific mapping. Section 3.2.3 discusses the feature-selection algorithm, in which features of Table 3.1 are normalized and mapped onto a simplified feature space. Section 3.2.4 discusses the classification process, in which a classifier is used to map the features onto the set of class labels $(0 - easy, 1 - complex)$.

3.2.3 Feature selection

Feature selection is an important step in reducing a feature set to a low dimensional space, thereby allowing for optimal performance of the classifier [Hei04].

Table 3.1: *Demographic and sonographic preoperative data collected from 337 patients and by 23 surgeons in the period between 2005-2008 at the academic hospital klinikum rechts der isar of the technical university of Munich*

Feature	Data value
Sex	Binary value (male, female)
Body mass index (BMI)	Index value in kg/m^3
Age	Value in years
Wall thickening gallbladder	Binary value: 1 if ($> 2mm$) else 0
Number and size of bile calculi	Ordinary 1: $calculi < 5mm$, ordinary 2: $5mm < calculi < 12$, ordinary 3: $calculi > 12mm$, or numerous smaller stones, which requires packaging of gallbladder with stones
Inflammation	Binary value
Previous surgeries in the upper abdomen	Binary value
Surgeon experience	Ordinary scale degrees $(3, 2, 1)$ depending on the total number of performed procedures
Surgeon	1-2-K binary coding: 23 different surgeons / 23 different binary features

The common assumption in pattern recognition is that any valid dataset is fundamentally low dimensional, even if it comes in a high-dimensional form. Moreover, the omission of feature selection requires the use of complex classifiers, which may waste considerable computational power in optimizing irrelevant corners of the high-dimensional space. The goal of feature selection is therefore to choose a subset $X_{s,n}$ of the complete set $X_{p,n} = \{x_{i,j}, i = 1, .., P \vee j = 1, ..., N\}$ of features such that the subset can predict the output $Y = \{y_i, i = 1, ..., N\}$ with accuracy comparable to, or better than, the performance of the complete input set X, and with significant reduction in the costs of computation [Tax08].

Feature selection involves two basic requirements: a *criterion function* for assessing the subset of features and a *search algorithm* for creating such a subset [Hei04]. The criterion function used is the *Mahalanobis distance*, defined as:

$$D_{maha-s} = (\mu_1 - \mu_2)^T (\frac{\sum_1 + \sum_2}{2})(\mu_1 - \mu_2)$$

The Mahalanobis criterion measures the distance between class densities for each feature. The distance measure is based on the distance between the two means μ_1 (feature values for Class 0) and μ_2 (feature values for Class 1), which should be large. Simultaneously, the ***covariance matrices*** $\sum_1$ and $\sum_2$ should be small, indicating a small spread around the means [Hei04]. For high classification

performance, the two class densities $(0, 1)$ should be far apart. Hence, significant features are characterized as having high Mahalanobis distance between the two classes.

Dimensionality reduction is performed using a *forward-search algorithm*; by evaluating each feature subset individually. Forward-feature selection starts with the single most significant feature and adds the next most informative features in an iterative greedy procedure. Its capacity to isolate efficient features is obvious. The forward technique has several drawbacks for cases involving large feature sizes or high interdependence between features. For our dataset, however, we expect that forward feature selection may generate optimal feature subsets, given that our pre-operative features are limited in both size and interdependency.

3.2.4 Binary classification problem

The basic problem addressed in this paper involves the binary classification of pre-operative data into two classes: *easy* (Class 0) and *complex* (Class 1) surgeries. The classification process has two stages: training and testing. Given a training set of N patients: $\chi = \{(x_i, y_i), i = 1, ..., N\}$, with $x_i \in \mathbb{R}^p$ are p-dimensional feature vectors, and $y \in \{0, 1\}$ are class labels. The goal of the classification process is to predict the class label of a given case within the test set [Tax08], given its feature vector.

A number of classifiers are considered for evaluation, including the LDC, the quadratic discriminant classifier (QDC), the Parzen classifier and the support vector machine (SVM) classifier. We trained and tested the classifiers with our dataset in order to identify the classifier that would yield the best classification performance for this dataset, according to the criterion described in Section 3.2.5. It was also necessary to know the number of samples required for the optimal training of the classifiers. The experiments in Section 3.3 elaborate on those points by analysing the classification errors, the AUCs: areas under the ROC(Receiver operating curve) and the learning curves of the classifiers.

3.2.5 Classification performance criterion

Two criteria were used to measure classification performance: the classification error and the area under an ROC curve (AUC). The classification error is often used for classifier evaluation through the straightforward counting of the number of misclassified records in a test set. For our binary classification problem, assume that a classifier f is trained and evaluated on a test set $\chi = \{(x_i, y_i), i = 1, ..., N\}$, with $x_i \in \mathbb{R}^p$ as the p-dimensional feature vectors from Table 3.1 and $y \in \{\omega^+, \omega^-\}$ as class labels indicating whether surgeries are easy or complex. The classification error is estimated by: $\epsilon = \frac{1}{N} \sum_{i=1}^{N} I(f(x) \neq y_i)$, where $I(.)$ is the indicator function that outputs 1 when the statement is true and 0 otherwise [Tax08]. One disadvantage to this measure is that it is sensitive to class priors [Tax08].

The AUC error is a natural criterion for measuring the classification performance of a classifier. In basic terms, it estimates the probability that a randomly selected positive (easy surgeries: 257 cases) is ranked before a randomly selected negative (complex surgeries: 80 cases). It is a widely used measure of ranking performance. It can be calculated by $E = 1 - AUC = 1 - \frac{1}{N^+N^-}\sum_{i=1}^{N^+}\sum_{j=1}^{N^-} I(f(x_i) > f(x_j))$, where N^+ and N^- refer to the number of objects from the positive and negative classes, respectively. For our dataset, where $N^+ = 257$ and $N^- = 80$, the AUC error remains a relative measure independent of those priors. Furthermore, the AUC tends to generate a more stable estimate of performance than does the classification error [Tax08].

3.3 Experimental validation

To address the problem of predicting surgical complexity, Section 3.3.1 focuses on the identification of the optimal classifier for the pre-operative dataset described in Table 3.1 and the number of samples required for optimal training. Section 3.3.2 aims to identify which subsets of features allow the prediction of complexity with accuracy comparable to, or better than, the complete set of features.

To answer these questions, we conducted both classification and feature selection. For the experiments described in Section 3.3.1 and 3.3.2, we used the statistical toolbox PRTools [Dui07b] to compare a set that is representative of state-of-the-art classifiers: LDC, QDC, Parzen and SVM classifiers. The feature-selection algorithm of Section 3.2.3 was also applied to the dataset described in Table 3.1 in order to derive the optimal feature set.

3.3.1 Classifier evaluation results

The experiments described in this section were intended to classify surgeries into two classes: *complex* and *easy* using features of Table 3.1. The subsets for both training and testing were randomly selected from the dataset described in Table 3.1 with equal prior probabilities and equal sample size.

The first experiment explored the classification performance of state-of-the-art classifiers using ten-fold cross validation. The data were divided into ten subsets of equal size. The classifiers were trained ten times, each time omitting one of the subsets from training, but using only the omitted subset to compute the required errors, as defined in Section 3.2.5. Figure 3.1 illustrates both the classification error and the AUC error using ten-fold cross validation for LDC, QDC, Parzen and SVM classifiers on the dataset described in Table 3.1.

The experimental results reported in Figure 3.1 suggest that the LDC, SVM and Parzen classifiers performed well for the dataset described in Table 3.1. Of all classifiers, the SVM classifier showed the best AUC error (0.16) and the second best classification error (0.17). On the other hand, the LDC classifier had a lower classification error (0.16) but a higher AUC error (0.28), which was even

higher than the QDC AUC error (0.26) and the Parzen AUC error (0.27). Both the Parzen and the QDC classifiers showed higher classification error (0.19) when compared to the LDC and SVM classifiers. With regard to classification error, the LDC classifier outperformed the SVM, Parzen and QDC classifiers, thus yielding the minimal classification error. Nonetheless, the AUC results demonstrate that the SVM classifier performed well for classification.

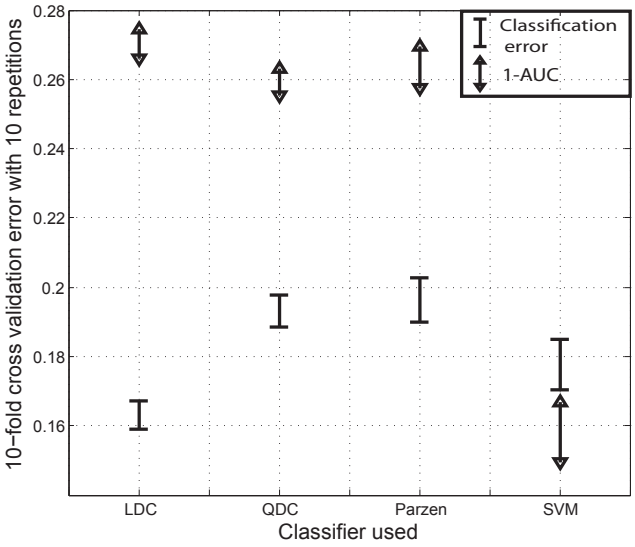


Figure 3.1: *Estimated classification and AUC error by tenfold cross validation for linear discriminant classifier LDC, Quadratic Discriminant Classifier QDC, Parzen and support vector machine Classifier SVM on the pre-operative dataset of Table 3.1*

The next experiment investigated the behaviour of the classifiers with regard to the verification of training-set sizes. Both the training set and the test set were randomly generated from the dataset described in Table 3.1, with equal prior probabilities but varying sample sizes. The resulting curve, the '*learning curve*', shows changes in the classification error for varying sizes. The learning curve indicates the classifier that is more suitable for small training set sizes and which has the most potential for performance improvement through the availability of more data. For this experiment as well, we used both the classification error and the AUC error.

Figure3.2 presents the learning curves of the LDC, QDC, Parzen and SVM classifiers on the dataset contained in Table 3.1, using the classification error. The size of the training set is a fraction of the total training size, ranging from 10% to 99% (80 patients) and limited by the size of the smallest class ($N^- = 80$). All classifiers fit the training set perfectly. A flat learning curve (as with the QDC) suggests that the classifier is already well trained, and more training data would

not help the classifier much. In contrast, a steeply decreasing learning curve (as with LDC, Parzen and SVM) suggests that better performance can be obtained through the availability of more training data. The learning-curve results using the classification error suggest the use of the LDC classifier, with a classification converging to 0.14.

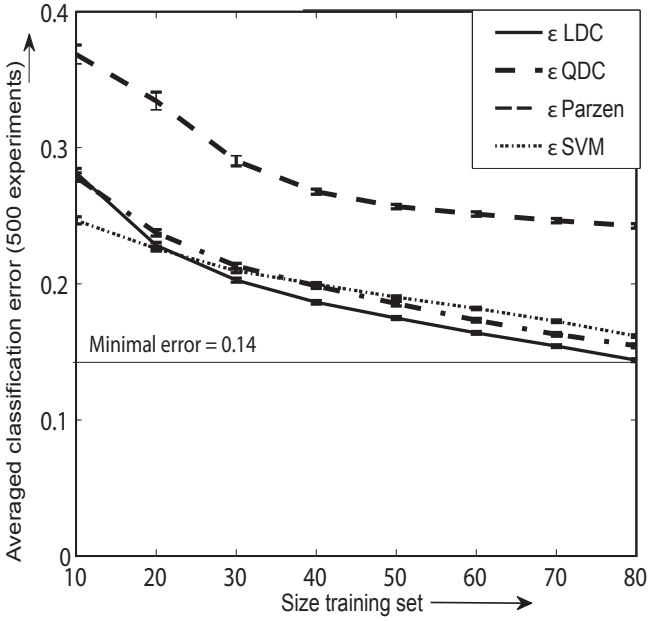


Figure 3.2: *Learning curve computed on the pre-operative dataset of Table 3.1 LDC, QDC, Parzen and SVM using the classification error*

Figure 3.3 illustrates the learning curves for the same classifiers on the same training and test sets, as specified in Figure 3.2. Instead of the classification error, however, the AUC error (1-AUC) is used. Notice that Figure 3.2 suggests that the LDC and SVM classifiers perform similarly for $N = 20$. As suggested by Figure 3.3, however, the SVM was clearly preferable for $N = 20$ in terms of AUC error. Note also that, as suggested in Figure 3.2, the LDC was the best classifier in terms of classification error and that the performance of the LDC, SVM and Parzen classifiers increased informally with increasing training size. As suggested in Figure 3.3, however, the SVM was the best classifier, and its performance in terms of AUC error did not improve significantly beyond $N = 40$. This suggests that the estimation of the SVM classifier was relatively reliable around $N = 40$.

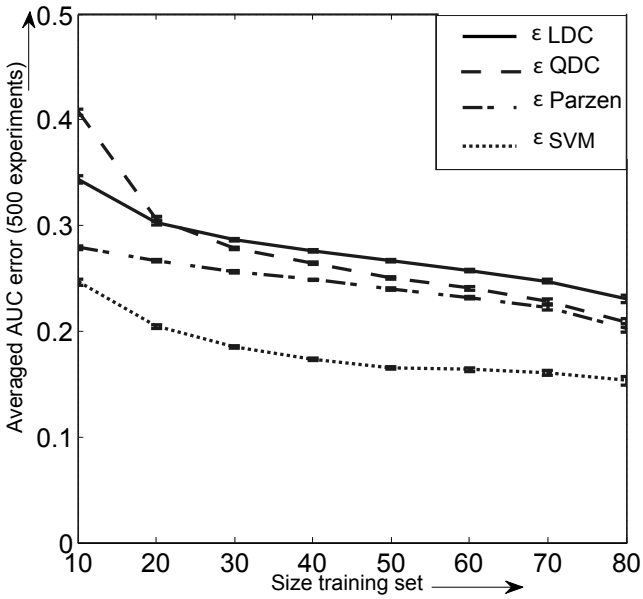


Figure 3.3: *Learning curve computed on the pre-operative dataset of Table 3.1 for LDC, QDC, Parzen and SVM using the AUC error*

3.3.2 Feature selection evaluation results

The experiments described in this section were intended to define the feature subset $X_{s,n}$ of the complete set $X_{p,n} = \{x_{i,j}, i = 1, .., P \vee j = 1, ..., N\}$ of features, such that the subset can predict the output $Y = \{y_i, i = 1, ..., N\}$ with an accuracy comparable to, or better than, the performance of the complete input set X. The feature-selection algorithm is explained in Section 3.2.3.

The first experiment was aimed at ranking the performance of individual features contained in Table 3.1. The goal was to rank a set of pre-operative features according to their contribution to the complexity of LAPCHOL surgery. Feature ranking is useful for determining the clinically relevant features from amongst all the available pre-operative parameters contained in Table 3.1. Ranking results can also provide insight into the stability of the feature-selection mechanism.

Figure 3.4 presents the ranking results for 200 experiments. For each experiment, 80% of the samples were randomly selected from the dataset and ranked according to the steps described in Section 3.2.3. Note that the same features could be ranked in different orders, depending upon the samples that were selected from the dataset. The general trend was for each rank level to be dominated by a small number of features. The main exception to this trend appeared in the higher levels (5 to 9), where the features were apparently less relevant for

the classification task. Note that the feature set was dominated by the surgeon from Level 5 through Level 9. This result reflects the effect of the labelling bias of the surgeon on the ranking results. Hence, the surgeon also influences the feature-selection mechanism for predicting operative complexity.

With regard to eliminating the effects of labelling bias on the results of feature ranking, the results shown in Figure 3.5 are similar to those shown in Figure 3.4 for only one surgeon with a high experience. Figures 3.4 and 3.5 show similar feature subsets up to Level 4. The bias is thus irrelevant for patients with inflammation, wall thickening, male sex and high BMI. These patients tend to be at high risk for complex LAPCHOL surgeries, regardless of labelling bias. The feature-selection mechanism is therefore considered stable for the subset of features including inflammation, wall thickening, sex and the BMI score.

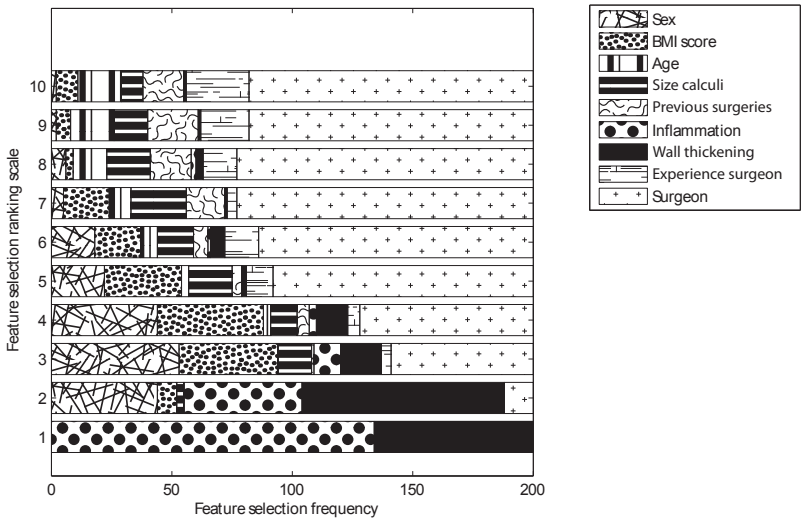


Figure 3.4: *Forward feature ranking for 200 experiments using the mahalanobis criteria function*

The next experiment used feature curves to investigate the relationship between the classification error and the dimensionality of the feature space. Learning curves typically report how large the training-set size should be in order to achieve optimal classifier performance. In contrast, feature curves provide information about the dimensionality of the feature space needed to achieve a low classification error. Figures 3.6 and 3.7 represent the feature curves of the LDC and SVM classifiers, respectively. Note that classification error was used as the criterion in these curves, with the result that the feature curve for the SVM classifier show

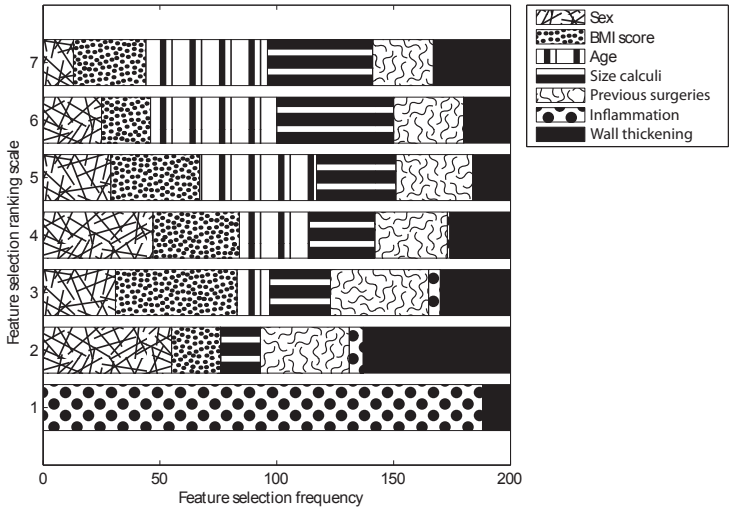


Figure 3.5: *Forward feature ranking for 200 experiments using the mahalanobis criteria function*

higher error rates than did that of the LDC. Both feature curves flatten out after the first four significant features. Moreover, the error increased with the use of features in addition to the four ranked features. As shown in Figure 3.6, the LDC achieved the lowest error rates when the first four ranked features were used. In contrast, Figure 3.7 shows that the SVM achieved its lowest error rates when the first three ranked features were used. We therefore conclude that the feature set derived from Figure 3.4 (i.e. inflammation, wall thickening, sex and BMI score) contains the optimal features for allowing the prediction of intra-operative complexity, with errors even lower those produced when the classifier is trained with the complete set of features contained in Table 3.1.

3.4 Discussion and conclusions

Intra-operative complexity can be predicted before surgery according to readily available pre-operative data. The problem addressed in this paper involves the classification of pre-operative data in two classes: *Easy* (Class 0) and *Complex* (Class 1) surgeries. Learning-curve results indicated a preference for the LDC classifier in terms of classification error, although the SVM classifier is preferable in terms of AUC error.

Feature selection was used to identify the pre-operative features that are

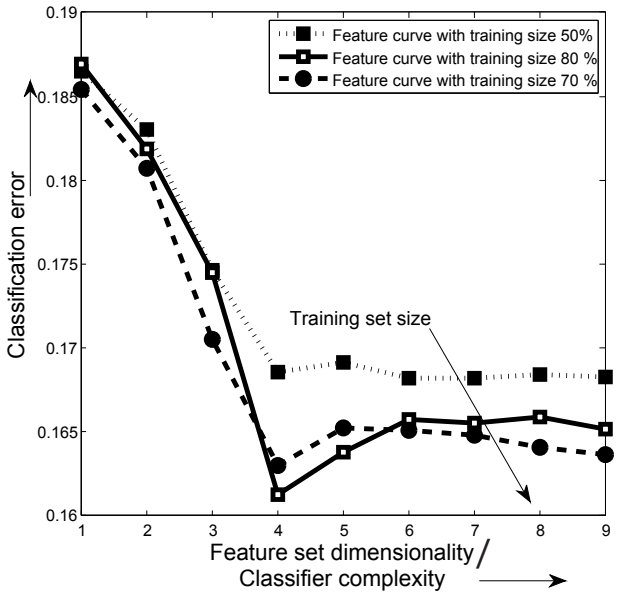


Figure 3.6: *Feature curve computed on the pre-operative dataset of Table 3.1 for LDC using the classification error*

relevant for predicting surgical complexity. The stability of these features was measured by repeating the selection experiment 200 times. The feature-selection mechanism proved stable for the subset of features including inflammation, wall thickening, sex and BMI score. Using learning curves, we demonstrated that these features are optimal for classifications using the LDC and SVM classifiers, thereby allowing the prediction of the complexity of surgeries. The performance achieved in these experiments was improved further when the classifier was trained with the complete set of pre-operative features.

Section 3.4.1 discusses the clinical relevance of the ranking results. Finally, Section 3.4.2 discusses directions for future work.

3.4.1 Conformity of ranking results with the clinical literature

Feature selection on pre-operative data identifies features that can be assessed preoperatively in order to determine operative complexity. Four pre-operative features were identified as relevant to the prediction of operative complexity. This section elaborates on the clinical relevance of the considered features and its conformity with our results.

The feature-ranking results revealed bias in the process of labelling surgical complexity. This indicates that surgeons have reasonable differences regarding the concept of surgical complexity, depending upon their level of experience, al-

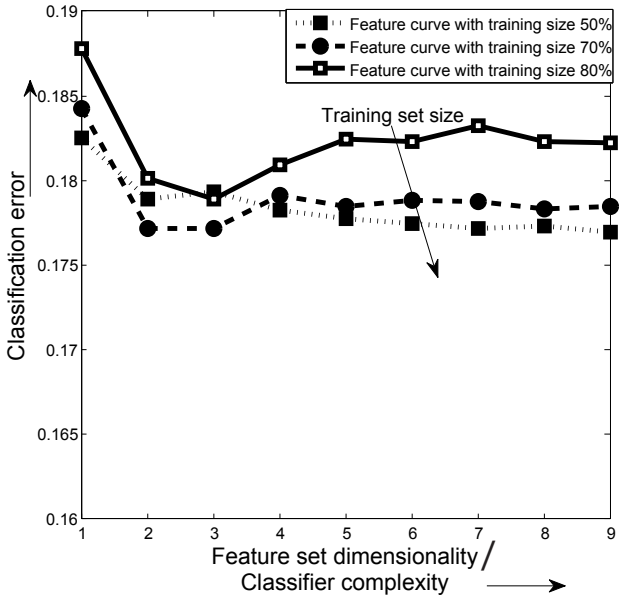


Figure 3.7: *Feature curve computed on the pre-operative dataset of Table 3.1 for support vector machine SVM classifier using the classification error*

though this bias is not relevant for the first four levels of the selection process. Based on the first four significant features in our dataset, we can conclude that subjectivity on the part of the surgeon when scoring complexity is reasonably low. We therefore consider the achieved classification results independent of the subjective opinions of surgeons regarding complexity.

The results of this study demonstrate that inflammation of the gallbladder is the most relevant complication factor for LAPCHOL procedures. This conforms to results from various clinical studies reporting that inflammation increases risks and complications during LAPCHOL surgery. One common risk involves the conversion from laparoscopic to open surgery. In a study involving 418 patients, Cox et al. [Cox93] showed that the frequency of conversion was 55.4% for patients with inflammation and 4% for patients with no inflammation. Another known risk is bile-duct injury (BDI), which is caused by the narrowing of the bile duct during the surgery. The narrow bile duct prevents the bile from draining, with the result that it backs up in the liver and spills over into the blood, causing obstructive jaundice. Based on a dataset of 2184 patients Georgiades et al. [Geo08] showed that the risk of BDI was 3.5 times higher in patients with inflammation, and they suggest that surgeons should not hesitate to convert to open surgery in the presence of inflammation. Our findings confirm that patients with inflammation tend to be at high risk for complex LAPCHOL surgery.

The other complexity factor identified in this paper is wall thickening. A thickened gallbladder wall makes it difficult for surgeons to detach the gallbladder from the liver bed. Dinkel et al. [Din00] considered wall thickening as the most relevant indicator of technical difficulties during LAPCHOL. In their dataset of 75 patients, 19 had sonograms revealing gallbladder wall thickening ($> 4mm$). Surgical preparation difficulties in 16 of these patients led to open surgery in four patients. Our results showed that, regardless the type of complication (technical or anatomical), wall thickening is the second most prominent factor (after inflammation) for complexity during LAPCHOL procedures.

In this article, the sex of the patient was also identified as a relevant factor for complex LAPCHOL surgeries. The clinical literature indicates that adhesions and obstacles in anatomical identification occur more frequently among male patients [Gab09]. Zisman et al. [Zis96] claim that the probability of conversion is five times greater in males than it is in females. The reason for higher conversion rates in male patients remains unexplained in recent clinical publications [Gab09]. Even though [Zis96] reports that some male patients have thickened gallbladder walls measuring above the upper limit of 3mm (3.4 + 1.5), the measurements gallbladder walls in female patients did not exceed the normal rage (2.6 + 1.3). Our results confirm that the complexity of LAPCHOL surgery is increased significantly for male patients.

The last significant pre-operative feature that our study identified as relevant for assessing intra-operative complexity is the BMI score. In the clinical literature, obesity is considered as a risk factor for laparoscopic procedures in general. For LAPCHOL, Gabriel et al. [Gab09] found that the highest percentage of conversion (28%) was observed in overweight patients. Our result confirm that, after inflammation, wall thickening and sex, BMI is the fourth relevant factor that contributes to the complexity of LAPCHOL surgeries.

According to our results, the complexity of LAPCHOL procedures seems to be less influenced by the patient's history of previous surgeries. This finding can be attributed to the advanced entry technique used by the surgeons in our dataset. This technique is known as the 'Hasson technique', in which the abdominal is incised under direct vision, in order to allow the insertion of the trocar. This risk of vascular complications associated with this technique is minimal, and even lower than that associated with the conventional blind approach [McK95]. Our results confirm that the use of the Hasson technique compensates for the history of abdominal surgery relative to the age of the patient, the history of previous surgeries in the upper abdomen and the size of the calculi. Moreover, the age of the patient apparently does not influence the course of surgery, although older patients do require more time than younger patients do for the induction of anaesthesia patients [Lew06]. Although longer aesthesia time is associated with prolonged surgical and recovery time, it is not associated with complex surgery. We also included the history of previous surgery in our dataset, as many of the complications associated with laparoscopic surgery arise from the creation of the abdominal entry (the pneumoperitoneum). Finally, the size of the calculi has

little influence on surgical complexity. Large gallstones may increase extraction time during surgery, thereby resulting in increased complexity. It is commonly believed that breaking large calculi inside the gallbladder in order to facilitate their removal could be dangerous. It is therefore customary to extract large calculi through large incisions, which facilitates removal in terms of both time and complexity [Ada96]. This practice also explains the low ranking of this feature in our dataset.

3.4.2 Future directions

This study shows that pre-operative data can be used to estimate surgical complexity preoperatively. In previous work [Bou10], we monitored the surgical workflow of LAPCHOL intraoperatively. In future studies, we aim to combine both pre-operative and intra-operative data in order to support both intra-operative and post-operative processes.

Chapter 4

Intraoperative: segmentation of workflow steps

Adapted from:
L.Bouarfa, P.P.Jonker and J.Dankelman
Published in the Journal of Biomedical Informatics, (2010)
under the title "*Automatic discovery of surgical workflow steps using hidden Markov models*"

L. Bouarfa, P.P. Jonker and J. Dankelman
Published in the proceedings of the 1st Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI), MICCAI London, (2009)
under te title"*Surgical context discovery by monitoring low-level activities in the OR*",

abstract

Detecting surgical high-level tasks during surgery is an important task for surgical workflow analysis. Surgical high-level task recognition is also a challenging task for context-aware applications because of the inherent uncertainty and the complexity of the surgical environment. In this paper we present a framework for recognizing high-level tasks from low-level noisy sensor data. Preliminary results, on a noiseless dataset of ten surgical procedures, shows that it is possible to recognize surgical high-level tasks with detection accuracies up to 90%. Introducing missed and ghost errors to the sensor data results in a significant decrease of the recognition accuracy. This supports our claim to use a cleaning algorithm before the training step.

4.1 Introduction

Recent years have seen a growing scientific and industrial interest in surgical workflow analysis [Neu06a; Blu08]. Workflow analysis is widely used in the domain of business process modeling (BPM) to improve organizational performance [Aal02]. Usually, the workflow follows a formal description of processes in the form of flow-diagrams, showing directed flows between the process steps. However, surgical tasks are hard to model with such a formal approach. Human abstraction in surgery is impossible, because of the more complex tasks performed by the surgeon, including aspects like cognition, uncertainty and skill [Sho06]. In this case, the formal approach needs to be adapted to deal with the complexity and uncertainty of the surgical environment.

To enable workflow modelling in surgery, surgical context information needs to be considered [Neu06a]. Therefore we need an abstraction layer that provides context data from sensing devices (e.g. sensors, video, etc.). Such a layer needs to infer surgical context information from a set or series of observations. This article aims to build upon already available context-aware techniques to recognize high-level surgical tasks using low-level information available in the OR. We first propose a conceptual framework to infer high-level tasks from low-level sensor data. We then attempt to answer the following questions: (1) how accurate can we predict high-level tasks using noise-free low-level instrument signals? and (2) how does the accuracy of the system respond to common sensor noise?

This paper is organized as follows: Section 2 gives background information. In Section 3 a conceptual framework for inferring high-level tasks from low-level sensor data is introduced. To evaluate the clarity and the reliability of the conceptual framework, ten surgical procedures are considered for the pilot study represented in Section 4. The experimental evaluation of our framework is discussed in Section 5. Related work is discussed in Section 6.Finally, Section 7 concludes this paper and give recommendations for future research.

4.2 Background

4.2.1 On inferring high-level tasks from low-level tasks

The objective is to infer a specific *high-level task* (HLT) from a set of observable *low-level tasks* (LLT). We considered that surgical workflow is composed of a number of high-level tasks $HLTs$. Although two $HLTs$ might be semantically identical, they can consist of different LLT sequences. We consider tasks performed by surgeons in the OR as $HLTs$ having the following characteristics:

- goal-oriented
- characterized by planning and manoeuvring protocols

- not described by a single (LLT) sequence, and thus may be performed in various ways

The mapping gap

The problem of mapping $LLTs$ to a specific HLT is known as the semantic gap. The semantics of a specific task depend on the context in which it is regarded. This requires transferring high-level tacit knowledge of human agents to explicit knowledge, a process known as articulation [Pat99]. The $LLTs$ that are most interesting for deducing the underlying $HLTs$ are those that:

- allow the discrimination over a large number of other LLT sets that correspond to other $HLTs$
- are invariant to task distortion, i.e. when a specific HLT is performed in different ways
- are compact in size. A small-set of $LLTs$ is beneficial for complexity constraints, since otherwise a large number of LLT-sets need to be stored and monitored in the environment. An excessively short representation, might not be sufficient to discriminate among similar $HLTs$
- are easy to monitor. The monitoring of $LLTs$ should not be complex. For real-time performance, the system requires high computational efficiency for both the monitoring of the LLT-sequence and the inference of the corresponding HLT

system parameters

The parameters of a HLT recognition system should be chosen based on the application and the cognitive environment or the context in which it is used. They are useful to evaluate and compare different HLT recognition systems. A number of these parameters include the following:

- *Robustness/ invariance :* ability to accurately infer a specific HLT regardless of the level of variance in task execution and the level of distortion in the environment (e.g. unexpected situations)
- *Discriminative power :* the ability to discriminate between similar, but different $HLTs$ (i.e. not the same). This may be conflicting with other requirements, such as robustness and complexity.
- *Accuracy :* the number of correct HLT inferences, missed inferences and wrong inferences

4.2.2 On the description of Laparoscopic Cholesystectomy

Laparoscopic cholystectomy (LapChol) is a highly standardized surgical procedure, where a patient's gallbladder is removed in case of inflammations. Cholecystectomy is performed under general anesthesia. Initially, a small needle is inserted into the peritoneal cavity for inflating the abdomen with carbon dioxide. This provides room for easier viewing and for the surgical manipulations to be performed.

To gain access to the gallbladder, four trocars are placed in the abdomen of a patient. For the laparoscopic camera is inserted in a 10 mm trocar (T1) placed at the umbilicus. The second 10 mm trocar is placed midway between the umbilicus and the xiphoid process, 2 to 3 cm on the left of the midline. The third 5 mm trocar (T3) is placed in the right iliac fossa. The fourth trocar (T4) is placed under the right costal margin to retract the liver or under the left costal margin to push the duodenum [Sli95].

First a retraction device is inserted in trocar T3. The right liver lobe is elevated. The laparoscopic camera is changed from trocar T2 to T1, to provide sufficient view of the surgical field. Finally, a grasping forceps is inserted into T4 and the dissection device in T2. The primary step of the surgical procedure is to dissect the area which includes the bile duct and the cystic artery (Calot's triangle). This is done by blunt dissection with a forceps and cutting current. In case of bleedings, coagulation current is used. If both structures are clearly visible, each of them is clipped with three clips, followed by cutting both structures between the clips with laparoscopic scissors. The following step is dissection of the gallbladder. In laparoscopic surgery, this is done by touching the areas between gallbladder and liver and applying cutting current. To remove the dissected gallbladder a salvage bag is inserted into the abdomen. The gallbladder is packed up into the bag and the bag is extracted together with trocar T1. In case of big stones, the bag cannot be extracted through the trocar incision. In that case, the calculi are extracted extra corporeally out of the salvage bag. Thus, the content of the bag is adequately reduced to pull it out.

Finally, the surgical area is explored one last time to detect and stop any bleedings. A drainage is inserted through a trocar hole and all instruments are removed. The trocars are extracted under visual control and the incisions are closed by sutures. During the procedure, in case of bleedings in the operation field, a device which allows flushing and suction is used. Also controlling for bleedings after extraction of the gallbladder is done with this device.

4.3 Conceptual Framework

To infer $HLTs$ from sensor data, we present an embedded framework in Figure 4.1. This framework allows for the cleaning of noisy sensor data by taking advantage of Bayesian Networks to infer the correct LLT from faulty sensor readings and fill

gaps in the dataset. The inferred $LLTs$ are further used to infer the corresponding $HLTs$ using HMMs. This system allows the inference of a specific HLT based not only on the available sensor data related to their $LLTs$, but also on their previously inferred $HLTs$.

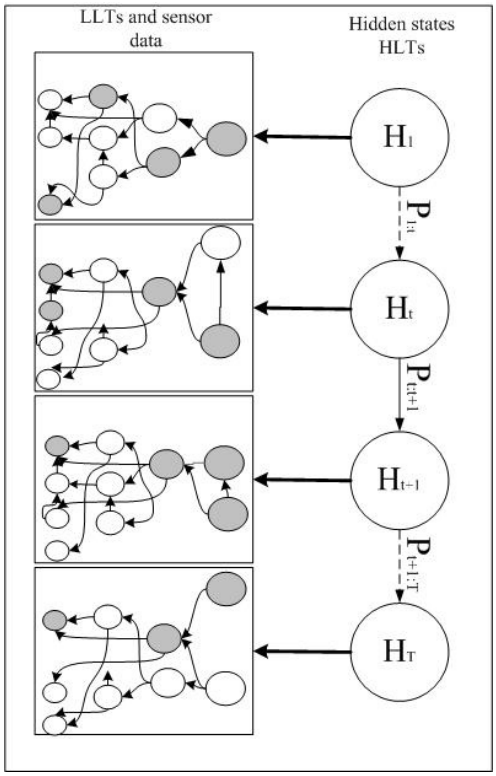


Figure 4.1: *Conceptual framework: embedded Bayesian Hidden Markov Model*

4.3.1 LLT-inference

To infer $LLTs$ from sensor data we need a classifier that takes as input, the sensor object (e.g. an RFID tag) $<< S, t > | < f_1, f_2, .., f_n >>$. Each f_i is a feature describing one characteristic of the object (e.g. tag) identified by the sensor S at time t, and makes a prediction of the form $<< S, t >: O, conf >$, where O is binary value, if $O = true$ the tracked object is detected, and $conf$ is the prediction confidence of the classifier [Gon07].

The process of inferring $LLTs$ is known as the cleaning process of sensor data [Dar07; Gon07]. Since sensor data is known to be noisy, a cleaning process assumes there is a hidden process that determines the true signal of the sensor,

such as presence of a tag in case of RFID, from a noisy and uncompleted set of features. To cope with the incomplete set of sensor signals, we propose to use Bayesian Networks to define the structure of the sensor signals that occur for a specific LLT. In case of RFID, features may describe one or more characteristics of the tag detected: the item to which the tag is attached, the location where the reading took place or the reader with which the tag is detected.

4.3.2 HLT-inference

To infer $HLTs$ from the observed $LLTs$ a classifier is used that takes as input the observed $LLTs$, $<< C, t >, < O_1, O_2, .., O_k >>$. Each O_i is the observation at time t, and makes a prediction of the form $<< C, t >: H, conf >$, where H is the value corresponding to the inferred HLT, and $conf$ is the prediction confidence of the classfier.

When the cognitive environment deals with a manoeuvring protocol, it is necessary to include knowledge from previous HLT inferences. DBN allows the representation of time constrained causality, i.e. when and if events occur and the periodic nature of processes. It is normally assumed that the model parameters (transition probabilities and model structure) of the temporal network do not change, i.e. the model is invariant [Jen01]. A special category of DBN is HMM. HMM is a strictly repetitive model with an extra assumption that the past has no impact on the future given the present [Jen01]. This means that the next HLT depends only on the current HLT. The $HLTs$ represent the possible hidden states of the HMM. The observable parameters of the HMM are the LLT nodes.

4.4 Pilot study

This pilot study was conducted to evaluate the clarity and reliability of the conceptual framework in recognizing surgical $HLTs$ using noise-free low level instrument signals. As such this study assumes perfect classification of sensor data, i.e. with $conf = 1$.

4.4.1 Dataset

In this pilot study ten LapChol procedures were recorded using three cameras. Three videos were recorded per procedure: an overview video of the OR, one pointing at the surgical toolbox and the laparoscopic video (See Figure 4.2). These video are synchronized and annotated using Elan software [Joh05].

Figure 4.2: *3 video streams (overview, surgical table, laparoscopic view)*

$HLT-$set

The $HLTs$ are the laparoscopic surgical steps as described in the hospital's LapChol protocol[1]. In total the five laparoscopic surgical steps are considered as $HLTs$, as illustrated in Table 4.1. Note that only laparoscopic $HLTs$ are considered, open surgical steps like incision and suturing are excluded. The definition of the surgical steps was verified with the co-operating surgeon[2] . For each sur-

1) skeletonization of calot's triangle
2) clipping and dissection
3) gallbladder removal
4) gallbladder packaging
5) cleaning

Table 4.1: *surgical steps of laparoscopic cholecystectomy*

gery HLT training samples are created from the annotated video as illustrated in Figure 4.3. The $HLTs$ are represented as discrete signals, each discrete level (from 1-5) corresponds to one of the five surgical steps of table 4.1. This results in the HLT**-signals** illustrated in Figure 4.3(a).

LLT-set

The $LLTs$ are represented as binary signals corresponding to instrument utilization; 1 if an instrument is in used, 0 if not in use. The resulting LLT**-signals** are displayed in Figure 4.3(b).

4.4.2 LLT pre-processing

The pre-processing step should retain the maximum contextual relevant information from the monitored $LLTs$. At this stage an invariant observation set $O = O_1, O_2, ..., O_n$ should be calculated from the observed $LLTs$. This observation set should allow the inference of similar $HLTs$ regardless of the level of

[1]LapChol protocol of Reinier de Graaf Hospital (RdGG), Delft, The Netherlands
[2]Dr. L.P.S. Stassen, a head surgeon at RdGG hospital and author of the LapChol protocol

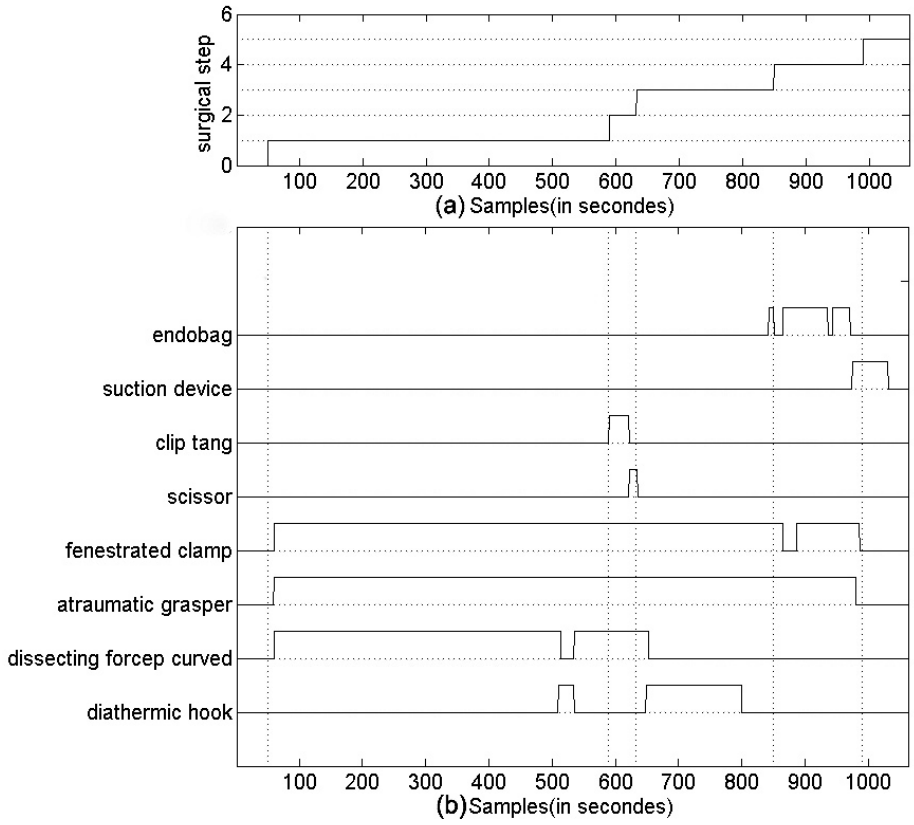


Figure 4.3: *Example of how training samples are created for (a) $HLT-$signals as a discrete levels of each surgical step(b) $LLT-$signals as binary signals corresponding to instrument utilization*

variance in their execution. This is a consequence of the robustness requirement from section 4.2.1. It should also allow the discrimination over a large number of other observation sets that correspond to other $HLTs$. Note that this requirement is conflicting with the invariance requirement. Both the robustness and the discriminating power are important for the evaluation of the system performance.

At this stage it is necessary to take the characteristics of the dataset in consideration. In the LapChol procedure four trocars are inserted to introduce the laparoscopic instruments in the patient's body. Figure 4.4 illustrates the use of the trocars.

- one main trocar (master), is maintained by the dominating hand of the surgeon. It is mainly used to insert instruments like dissectors, scissors, to remove the gallbladder

- two other trocars (slaves), are maintained by the non-dominant hand of the surgeon and highly correlated to the master trocar. They are mainly used to insert gaspers to hold the gallbladder for removal
- one view trocar is used to insert the laparoscopic camera

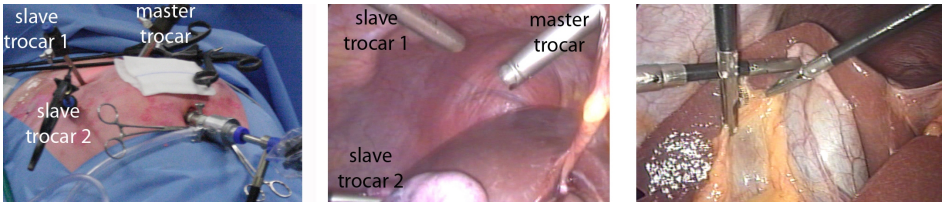


Figure 4.4: *Trocars inputs in Laparoscopic Cholestoctomy*

In total, 10 instruments are used during the LapChol procedure, from which 3 can be used simultaneously, leading to $\begin{pmatrix} 10 \\ 3 \end{pmatrix}$ possible sets. To reduce the observation set for training we consider two datasets:

- the first dataset is pre-processed for the LLT "taking instrument X from the surgical toolbox". The LLT-observation matrix is converted to the observation-set $O_{toolbox} = O_1, O_2, .., O_k$ with O being the label of the last changed instrument value (both 0 and 1 are considered)
- the second dataset is pre-processed for the LLT "inserting instrument X into the master trocar". The LLT-observation matrix is converted to the observation-set $O_{trocar} = O_1, O_2, ..., O_n$, with O being the label of the instrument inserted into the master trocar. Note that this LLT exploits the high correlation between the slave and the master trocars.

In prior work, Padoy et al. [Pad08] used a dataset similar to $O_{toolbox}$, the major difference is that we adopt an asynchronous processing by excluding the time component in both datasets, by using the label of the last used instrument. Instead, in [Pad08] they include all sample points in the training $O_{k,t} = 1$ if instrument k is active at time t. In previous work [Bou09] we demonstrated that the asynchronous approach of training outperforms the synchronous approach.

4.4.3 HMM Training

A HMM can be denoted as follows, $\lambda = (\pi, A, B)$ where π described the initial distribution, A is the transition matrix of the Markov process and $B = b_i(x)$ is the emission matrix, indicting the probability of emission of symbol x from a hidden state i.

Training a HMM consists of estimating the transition matrix A and the emission matrix B according the observed sequences. Here, the Baum-Welch EM algorithm is used on both observation-sets $O_{toolbox}$ and O_{trocar}. The Baum-Welch algorithm estimates model parameters (A and B) from the observed sequence while maximizing the log-likelihood of the model. For inference of the surgical steps, the Viterbi algorithm is used to calculate the most likely path of states (the sequence of visited states), also called the Viterbi path. The Viterbi path relies on global criteria, meaning that the low-level variation in the data is smoothed. The use of Viterbi path reduces false rejections in the system.

HMM-Outputs

Given an observation-set $O = O_1, O_2, .., O_n$ and a HMM, the most probable state sequence is found using the Viterbi algorithm. This sequence was found for all the ten cases using the trained HMM. Figure 4.5 shows an example of an HMM output using the LLT-dataset O_{trocar} of a specific LapChol-procedure. The inferred

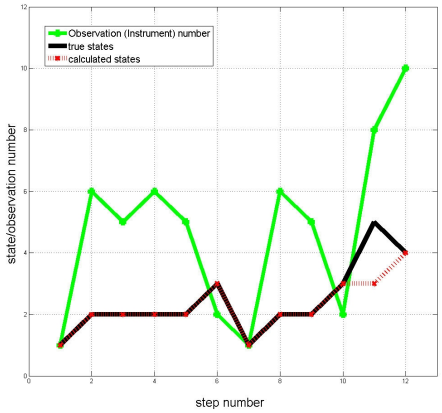


Figure 4.5: *HMM outputs: True and inferred measurements*

states $HLTs$ by the HMM (red line) matches the true states of the system (black-line) for the majority of the data-points. Moreover, the inferred states are more sensitive to instrument transitions than the subjective ground truth states as defined by the surgeon. In the next section, the performance of the system is evaluated on both datasets in more detail.

4.5 Experimental Results

In this section we aim to answer several open questions that have not been addressed in previous research: (1) How accurate can we predict $HLTs$ using noise-

free instrument sensor data? and (2) How does the accuracy of the system respond to common sensor noise?

4.5.1 How accurate can we predict $HLTs$ using noise-free instrument sensor data?

In this section we evaluate the performance of the HMM at predicting surgical $HLTs$ using the LLT-datasets described in Section 4.4.1. We performed a full cross validation. Within a group of ten observations one set is used for validation and the remaining nine sets are used for training. This is performed on each possible combination of training and validation sets, and thus for a total of ten times.

To evaluate the accuracy of the system in recognizing surgical steps, the *frequency error rate (FER)* is calculated for each state as the percentage of time that the surgical $HLTs$ are correctly detected. Figure 4.6 illustrates the *FER* results for both datasets. The system shows a total accuracy of 90% of detected states. The result demonstrates that data from the master trocar alone lead to a more robust inference mechanism for training. As expected, activities with higher accuracies were generally those with more data points (i.e. more instruments used). For the LapChol procedure, they were the "clipping" and the "removal" steps. The lower performance of the "cleaning" step can be attributed to the relatively few samples of this HLT in our dataset. Moreover, some instruments are highly robust compared to others in indicating their corresponding surgical step. For example, the "clip tang" is used only in the "clipping" step, and provides a robust and highly discriminative indication of this HLT. In this case, the HMM is analogous to a weighted voting mechanism.

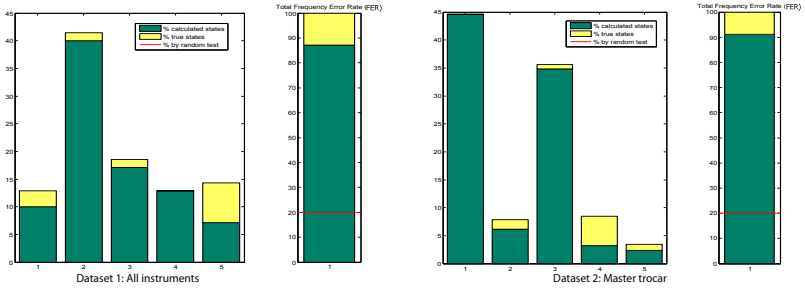


Figure 4.6: *frequency error rate (FER) of surgical HLT detection for, the red line is the level of a random test.*

Figure 4.7 shows the confusion matrices, whose row and column index indicate the recognized and ground-truth surgical steps, respectively. Each element a_{ij} in the confusion matrix indicates the percentage of data points O_n from a particular

hidden state j that are assigned to another hidden state i by the HMM classifier. E.g. the value "43" in the upper left cell of the matrix indicates that 43% of the data points O_n from the hidden state 1 are assigned to the hidden state 1 by the HMM classifier. Figure 4.7 also shows the precision and recall data for both confusion matrices, which are defined as follows:

$$Precision = \frac{tp}{tp + fp} \tag{4.1}$$

$$Recall = \frac{tp}{tp + fn} \tag{4.2}$$

where tp is the number of positive samples in the input data that have been correctly identified by the HMM classifier; fp is the number of negative samples that have been incorrectly identified as positive by the classifier; fn is the number of positive samples that have been identified as negative by the classifier.

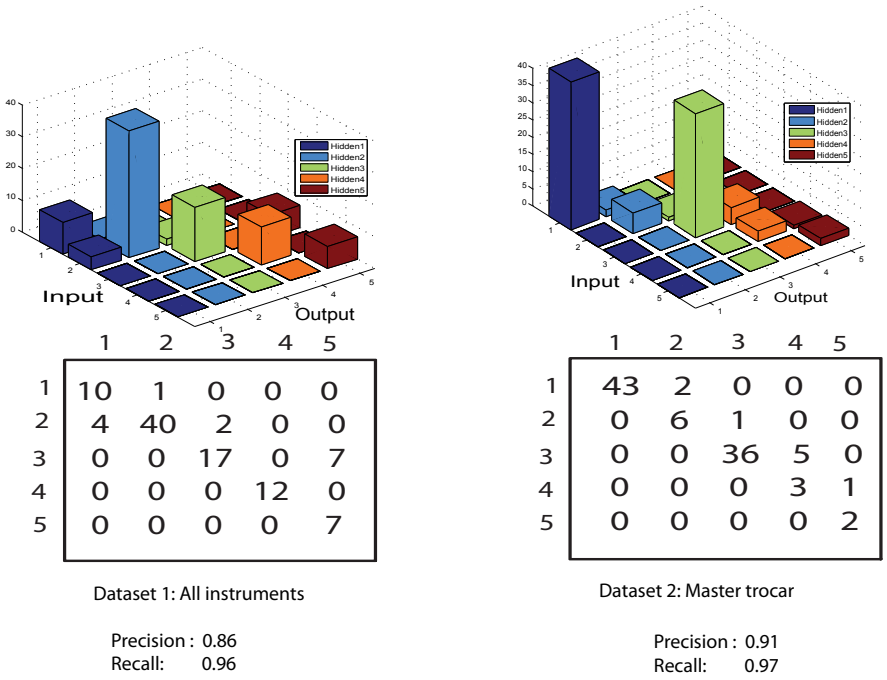


Figure 4.7: *Confusion matrix, precision and recall*

As shown in Figure 4.7, good results can be obtained from training with both datasets. Cases for which there are a few samples tend to have poorer performance. For example, for Dataset 1, Step 5 and 3 were confused, which

results in an error rate of 50%(4.6, yellow area, Step 5, Dataset1). However, in Dataset 2, Step 4 is classified with an accuracy of 37% (4.6, green area, Step 4, Dataset 2), because of high confusion with Step 3. These errors can be attributed to the relatively low number of training and test samples.

Considering the recall and precision characteristics, both datasets show good results in a recall test. Accordingly, a state can be correctly inferred with high probability (i.e. high robustness). However, recall alone is not sufficient, as the number of wrong inferences should be as low as possible as well (i.e. high precision). The precision values show that training with Dataset 1 results in more false positives than Dataset 2. Hence, those categories with more samples (data points), will result in an increase in false classifications, causing the HMM classifier to misclassify new data. The conclusion regarding Figure 4.7 is that data from the master trocar alone (Dataset 2) result in more discriminative power for the HMM classifier.

To measure the similarity between the estimated HMMs, the Kullback-Leibler Distance (KLD) is measured between each pair of Markov models λ_1 and λ_2. The KLD is widely used as a distance measure between HMMs [Zen09]. The KLD is computed in the literature using the Monte-Carlo approach as follows:

$$d(\lambda_1, \lambda_2) \approx (1/T) * (log(p(O_1|\lambda_1)) - (log(p(O_1|\lambda_2)))) \tag{4.3}$$

where, O_1 is a sequence generated by model λ_1, and T is the sequence length. In case of a stationary HMM ($\pi_t = \pi_{t-1} = \pi_s$), for a given sequence $O = o_1, o_2...o_T$

$$P(O|\lambda) = Pr(o_1)Pr(o_2)...Pr(o_T) \tag{4.4}$$

where, Pr is the output distribution of HMM, and can be calculated as follows:

$$Pr(x) = \sum_{i=1}^{N} (\pi_{s,i}) * b_i(x) \tag{4.5}$$

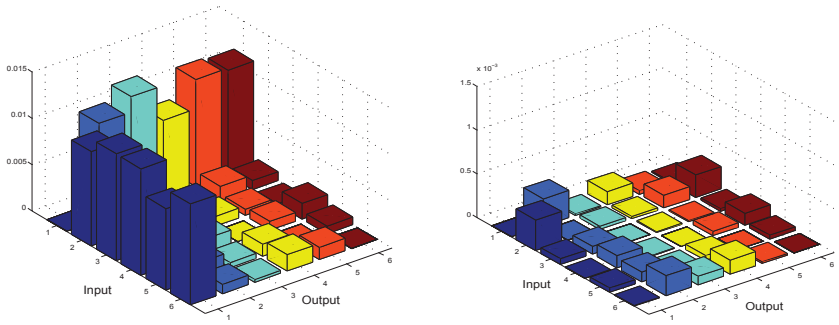
where, N is the number of hidden states of HMM, $\pi_{s,i}$ is the stationary probability of state i, and $b_i(x)$ is the emission probability of symbol x from a hidden state i. Hence, Eqn. 7.1 results in:

$$\begin{aligned} d(\lambda_1, \lambda_2) &\approx (1/T) * ((\sum_{i=1}^{T} logPr_1(O_i)) - (\sum_{i=1}^{T} logPr_2(O_i))) \\ &= (1/T) * ((\sum_{i=1}^{T} (logPr_1(O_i)) - logPr_2(O_i))) \end{aligned}$$

Where $T = 5000$, further the KLD is computed with the symmetric version as follows:

$$d_{KL}(\lambda_1, \lambda_2) = 1/2 * (d(\lambda_1, \lambda_2) + d(\lambda_2, \lambda_1)) \tag{4.6}$$

Figure 4.8 shows a matrix plot of the KLD measured between all possible pairs of 6 HMMs. Each element $d_{KL}(i, j)$ in the KLD matrix indicates the KLD between the pair HMMs λ_i and λ_j as defined in Eqn. 4.6. We can see that the KLD between HMMs trained with the trocar dataset is smaller compared to Dataset 1 that includes all surgical instruments. This confirms the results from the evaluation of the recall and precision metrics that data from the master trocar alone (Dataset 2) result in more discriminative power for the trained HMM classifier.



Dataset1: All instrumens
mean: 24 e-6
variance: 3360.2 e-8

Dataset2: Master trocar
mean: 8.3 e-6
variance: 1.4 e-8

Figure 4.8: *Kullback Leibler Distance between all possible pair of 6 HMMs*

4.5.2 How does the accuracy of the system respond to common sensor noise?

Surgical instruments can be monitored using different kind of sensors. We consider the use of state-change sensors like RFID tags to monitor the O_{trocar} signals. These sensors allow easy and continuous data collection, however they suffer from two main types of noise [Eng05]:

- Missed tag errors: results in no data, such as the identifier stored on the tag, being collected from a tag by a specific tag reader.
- Ghost tag read errors: results in erroneous data, specifically an identifier that is not stored on any tag within the reader's field, being "read" by a reader and reported as correct data.

These two errors are simulated by increasing the missed rate and the ghost-error rate of cross test-sets O_{trocar}. Further, a well trained HMM is used to test the noisy sets. Figures 4.9, and 4.10 show the result of increasing the missed rate and the ghost rate from 10% to 90%. Both errors results in significant degradation of

recognition accuracy. Introducing ghost errors result in a linear decrease of the recognition accuracy, while missed readings result in a non-linear degradation of recognition accuracy. The non-linearity of the latter error is due to the significance of the missed instrument in indicating the corresponding surgical step.

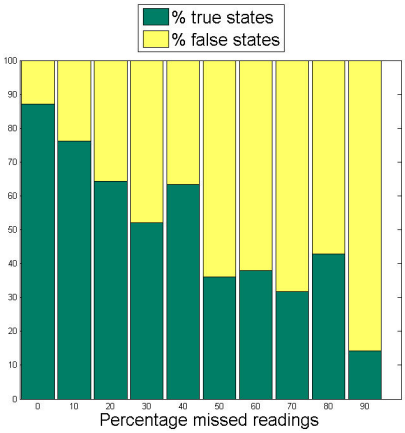


Figure 4.9: *Missed tag errors*

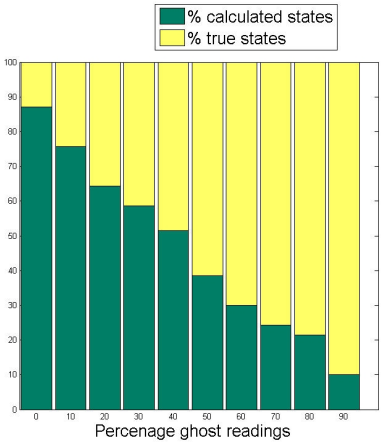


Figure 4.10: *Ghost tag errors*

4.6 Related work

In context-aware applications, small and simple state-change sensors are used for data collection. Vankipuram et al. [Vak09] used active RFID tags in a dynamic medical environment for human and equipment tracking. Tapia et al. [Tap04] used "tape on and forget" sensors to recognize activities in home setting, and Pham et al. [Pha07] used ultrasonic sensors to classify trajectories of movement of patients and elderly in indoor environments. Besides ubiquitous computing, human activity recognition is studied in vision and media research. In the past two decades, significant progress has been made in specific areas such as speech recognition, face recognition and video surveillance [Rab89; Pan07].

Sensors allow continuous data collection on a large scale. However, there are various problems which hinder the adoption of sensors in reliability critical environments, such as noisy sensor outputs, missed readings and inferences. Accordingly, different data cleaning approaches are proposed in literature to allow correct interpretation and analysis of sensor data. Darcy et al. [Dar07] improved the missed data restoration process of RFID tags using Bayesian Networks. Vankipuram et al. [Vak09] found that the tag data is extremely noisy and used Hidden Markov Models to improve the motion recognition accuracy. Gonzalez et al. [Gon07] proposed a Dynamic Bayesian Networks (DBN) based cleaning method of RFID data sets that takes tag readings as noisy observations of hidden states and performs effective data cleaning.

To infer $HLTs$ from sensor data, graphical probabilistic models are used with the underlying assumption that there exist hidden states that represent the $HLTs$, and that the hidden states are evolving. Graphical probabilistic models enable the inference of hidden states from the observable $LLTs$ up to temporal or causal relationships, for example Bayesian Networks (BN) [Tap04; Gon07], Hidden Markov Models (HMM) [Pad08] and Conditional Random Field models (CRF) [Hu08]. For pre-selection of the observation set, Tapia et al. [Tap04; Gon07] used a feature window per HLT by assuming that different $HLTs$ have a different mean of their length in time (duration). The features used for inference are then calculated within the window size. This assumption is however not applicable to cognitive environments with high time-variability of $HLTs$. Hu et al. [Hu08] used an adapted version of the Conditional Random Field model to identify multiple-goal behaviours, such as concurrent and interleaving activities.

Recognition of surgical events has been adressed in [Lo07; Lo03] visual cues from endoscopic images for activity profilying is laparoscopic surgery. Blum et al. [Blu10] used visual cues from laparoscopic video to segment surgical steps using canonical correlation analysis. In [Jam07] information from an eye-gaze tracking system is used to detect the clipping step of cholecystectomy. Aproaches like Dynamic Time Warping are also used for segmenting surgical steps of surgery using laparoscopic tool usage [Ahm06] and [Pad07]. Padoy et al. [Pad08; Pad10] also used instruments signals to infer surgical $HLTs$. The signals were directly processed by the inference engine in the form of a Hidden Markov Model (HMM).

The fact that no pre-processing was used to filter robust $LLTs$, required extra filtering in the inference phase by merging states of the HMM.

Instruments are valuable signals in the OR environment for surgical activity recognition, as it is easy to monitor whether or not an instrument is in use, by using RFID tags or applying image processing algorithms on laparoscopic video. For an accurate recognition of $HLTs$, this paper proposes an integrated framework for inferring $LLTs$ and $HLTs$ from sensor data. The proposed framework is used to show how accurately instrument signals can predict surgical $HLTs$, how these signals can be pre-processed to obtain a more robust and discriminative observation sequence for training, and how to use RFID tags and image processing algorithms for an accurate and robust HLT recognition.

4.7 Discussion

In this paper, we proposed a framework to allow the inference of a specific high-level task based not only on the available sensor data, but also on their previously inferred high-level tasks. We posed two fundamental questions: (1) How accurate can we predict high-level tasks using noise-free instrument sensor data? and (2): How does the accuracy of the system respond to common sensor noise? By performing experiments with data based on ten laparoscopic cholecystectomy procedures, we showed that the system can predict 90% of the surgical high-level tasks using noise-free instrument sensor data.

In this framework, we proposed to take advantage of Bayesian Networks to clean noisy values of sensor readings and infer correct low-level task from faulty sensor readings. As we did not have real-sensor data available for testing, we empirically simulated the recognition accuracy by introducing missed readings and ghost readings rating from 10 to 90% in our training-set. Both errors results in significant degradation of recognition accuracy. This supports our claim to use a cleaning algorithm before the training step.

In addition to the cleaning and the inference algorithm, the preprocessing of sensor data is a crucial step. The preprocessing step should retain the maximum relevant domain information from sensor data and reduce the number of possible observations for training. Hence, we have demonstrated for the laparoscopic cholestoctomy procedure that sensor data from the master trocar leads to a more robust accurate and discriminative recognition compared to sensor data from the surgical toolbox.

In our pilot, we used instrument signals as data for automatic task recognition. This leads us to the question: can other sensor-friendly data be extracted from the OR for high-level task recognition? In a real operating room, a number of activities are performed by nurses, surgeons and surgeon-assistants. If the (high-level) tasks of every staff-member could be automatically recognized and the hierarchical relationships between these tasks could be derived, the output can provide crucial additional information on the overall surgical workflow.

In our experiments, we have shown that different recognition accuracies can be achieved under different levels of sensor noise. Further validation of automatically inferring the high-level tasks from real incomplete and/or noisy set of sensor data using the proposed Bayesian cleaning algorithm is desired. This issue is also related to the granularity requirement described in section 4.2.1. Hu et al. [Hu08] showed that different recognition accuracies can be achieved under different levels of granularities. For our application, it is very interesting to automatically set the task granularity level from available (incomplete or noisy) sensor data. This is important for applications in reliability environments, such as the OR, where a certain accuracy of recognition need to be achieved before system intrusion. The future challenge is to automatically set the level of granularity the system can support with the available (noisy) sensor data, given the hard constraint of high accuracy.

Chapter 5

Intraoperative: tracking of surgical instruments

L.Bouarfa, O.Akman, A. Schneider, P.P.Jonker, J.Dankelman
Published in the Journal of Artificial Intelligence in Medicine, (2011)
under the title "*In-vivo real-time tracking of surgical instrument in endoscopic video* "

abstract

Tracking instruments during surgery is becoming a useful acquisition tool for different applications. This article presents a tracking system to detect and track instruments in endoscopic video using biocompatible colour markers. The system tracks single or multiple instruments in the video. The originality of this method is that it combines continuously adaptive shift algorithm with Kalman-filter for real-time tracking of single and multiple surgical instruments during surgery. Preliminary results show that the proposed method has a real-time performance. Moreover it is robust to partial occlusion and smoke. The system shows high sensitivity and specificity results for blue, green and yellow colours. The achieved sensitivity and specificity results are sufficient to apply the system for real-time automatic recoding of surgical workflow in-vivo during surgery.

5.1 Introduction

Laparoscopy is a surgical technique in which real-time imaging is used during the surgery. While laparoscopic surgery produces more and more videos, there is limited work being done on extracting visual features for task automation. Yet this would be a major clinical added value, since it would allow for automatic measurement and assessment of surgical tasks.

Tracking surgical instruments offers interesting possibilities for different applications in surgery. In visual servoing applications, tracking surgical instruments is used to guide robotic arms using visual feedback from a camera system. A common scenario for visual servoing is the automatic guidance of the endoscopic camera by an assistant robot arm. For visual servoing applications, it is enough to retrieve the position of the tip of each instrument in the image; the camera can then be centred on an instrument. Another application is image-guided surgery (IGS) where optical tracking is commonly used to predict the position of instrument tip in the patient body by tracking instrument markers on the outside of the body. Another application of IGS is tracking marked instruments with ultrasound scans allowing visual feedback of instruments' position. During laparoscopic surgery, surgeons' gain most information, necessary to perform surgery, from the visual feedback of cameras, there is, however, limited amount of research performed on utilizing this visual information to guide autonomous information systems.

A new generation surgical information systems commonly includes measurement and assessment of surgical workflow. Measuring surgical workflow in laparoscopic surgery has been proven feasible by tracking surgical instruments [Bou10]. This would be of major clinical added value since it would allow to easily document and index surgeries. Furthermore, tracking surgical instruments is shown, in a lab setting, to be valuable to measure the surgical gestures and skills of surgeons [Meg06a]. This article presents a tracking system to detect and track instruments in laparoscopic video using biocompatible colour markers. It tracks single or multiple instruments using the endoscopic video feed. The originality of this method is the use of a continuously adaptive mean-shift algorithm, together with Kalman-filtering, allowing for real-time, simultaneous tracking of multiple surgical instruments during surgery.

5.2 Related work

Previous work on instrument tracking is mainly related to visual servoing and image-guided applications. Staub et al. [Sta10] proposes a switching servoing scheme, using both position and image based servoing to drive the instrument into the field view of the camera and to allow autonomous high precision positioning of surgical instruments in a complex setup with four robots. By tracking the instrument tip, Beasley et al. [Bea09] improved the motion accuracy of an existing image-guided tele-operation scheme involving flexing instruments. Stoll et al. [Sto06] used 3-dimensional ultrasound for visual servoing to guide a surgical instrument to a tracked target location. Wang et al. [Wan09]) used image recognition for automated inspection and identification of surgical instruments on the surgical table. Three types of surgical instrument of different size could be inspected by the system. Research from Rivera et al. [Riv08] utilizes RFID (Radio Frequency Identification) technology to aid in counting for all items used during surgery. With the design of the Scrub Nurse Robot (SNR) system, which is meant to replace a skilled human scrub nurse, Miyawaki et al. [Miy09] have developed an automatic acquisition system of surgical-instrument information for laparoscopic surgery by using RFID technology.

Tracking the position of instruments in endoscopic video is shown to be an important improvement for new generation surgical information systems. Real-time monitoring of instruments usage is shown to be effective in segmenting surgical workflow. In previous work [Bou10] we showed that surgical workflow activities can be detected with high accuracies using a time-indexed dataset of instrument utilization in a binary form (bit 1: in use, bit 0: not in use). Instrument binary signals are used for training a Hidden Markov Model (HMM) to infer the corresponding surgical activities. The surgical activities are detected using instrument signals with detection accuracies up to 90% [Bou10]. Another interesting application of instrument tracking is related to objective assessment of surgical skills. Megali et al. [Meg06a; Meg06b] presented a method that detects surgical gestures from kinematic data describing movements of surgical instruments in a simulated setting. The defined gestures are measured from experts in simulated setting and used to train a HMM [Meg06a; Meg06b]. The HMM is used as an expert model to evaluate the performance of surgeons with different abilities objectively. The model can efficiently be used to quantitively assess the surgical ability and to discriminate between experienced and novice surgeons. Thus far, surgical workflow and surgical gesture research can be performed only in a simulated setting. The big challenge is to measure both surgical workflow and surgical skills in-vivo during real surgical practice.

In this paper we present an adequate real-time instrument tracking tool that allows not only the detection of arbitrary instruments in endoscopic video but also distinguishes the different instruments in this same video by exploiting the preinstalled markers. Unfortunately, the natural visual features of the instrument are not utilized since they are not discriminative enough to distinguish the

different instruments in the current setup. Therefore, we propose the use of a biocompatible colour marker on the instruments and show that with this marker, we achieved fast, robust and discriminative instrument detection. We expect the results of this study to allow workflow segmentation real-time during surgery, to bring surgical skill assessment into the real surgical practice and to improve existing visual servoing and image-guided systems allowing the discrimination of different instruments and simultaneous multiple instrument tracking.

5.3 Method

5.3.1 Overview

The system architecture described in this paper is illustrated in figure 5.1 and the building blocks of the system are explained in detail in the following sections. Initially, the colour markers placed on the instruments are segmented from the background using colour information. Afterwards, the markers are tracked in the segmented regions to extract their trajectories.

5.3.2 Marker Segmentation

Segmentation is an important step for tracking in cluttered and occluded environments. Correct segmentation eliminates background regions and therefore improves the accuracy of the system (by providing to the rest of the system only the relevant foreground regions). "Foreground regions" can be defined as the regions that the markers can be located. The foreground regions can be detected by using the colour information available in the scene. Instruments can be segmented using colour information.

Initially, the colour markers are selected by using a HSV (hue, saturation, and value) colour map, such that they are positioned with maximum distance relative to each other in the hue space. This maximizes the hue difference between the markers and decreases possible confusion between the different markers.

A few images of a marker are captured under various illumination conditions and marker regions are segmented manually from the background. Different parts of the operation video are used to obtain different illumination conditions. Afterwards, these regions are utilized to build a marker-colour histogram S in hue-saturation colour space. For a given pixel m_i with hue-saturation values (h_i, s_i), the probability that m_i is a pixel from a marker region is

$$P(marker|m_i) = \frac{S(h_i, s_i)}{\sum_{k,l} S(h_k, s_l)} \tag{5.1}$$

The marker-colour histogram is updated by the new image histogram if $P(skin|(h, s))$ for a histogram bin (h, s) is greater than a predefined threshold

$$S_{t+1}(h, s) = wS_t(h, s) + (1 - w)S_{new}(h, s) \tag{5.2}$$

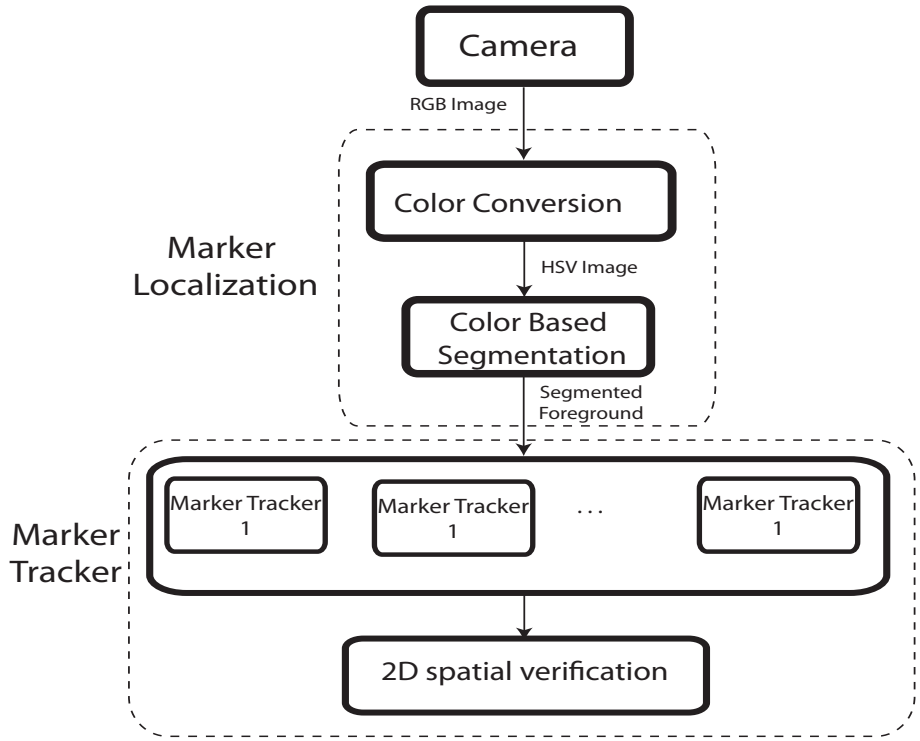


Figure 5.1: *System architecture*

Finally, a binary mask, M_c, is created by thresholding the pixel probabilities $P(skin|m_i)$. Morphological closing (dilation) is performed on the resulting mask, M_c to fill possible gaps in the masks and to extend the marker search region for robustness.

5.3.3 Instrument tracking via markers

The tracking system is schematically presented in figure 5.2. As a first step blob detection (connected component labelling) is performed on the mask, built in the previous Marker Segmentation step, M_c, to group foreground pixels. This helps to eliminate too small or too large regions that can be the result of camera noise. Afterwards, the markers that are positioned inside the blobs are tracked using the OpenCV implementation of a Continuously Adaptive Mean Shift (CAMShift) algorithm [ope]. This algorithm combines the basic mean shift algorithm [Com02] with an adaptive region sizing step. Markers can have different 2D sizes in the

video, depending on their 3D motion with respect to the user. The adaptive region sizing step of the algorithm can cope with the limitations originated from the varying sizes. A separate tracker for each marker is used together with its colour model. Also the models are continuously updated as long as the marker is detected to increase robustness against changing lighting conditions.

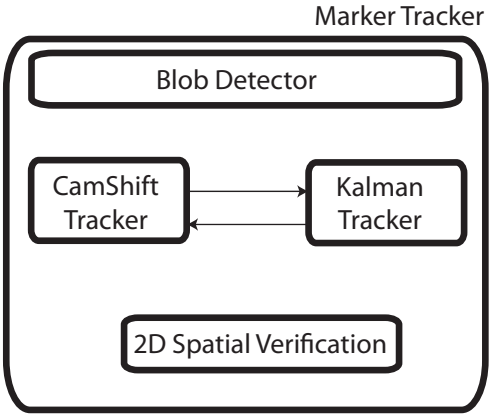


Figure 5.2: *Tracker overview*

Finally, a Kalman filter [Wel95] is employed together with the CAMShift tracker in order to cope with situations when the markers are not visible, lost due to the image noise or occluded by other instruments or abdominal tissues. The filters are updated with the measurements (detected 2D positions of the markers) from the CAMShift tracker. When the CAMShift tracker fails to track markers, Kalman predictions are used as the 2D positions of the markers for n frames. If the CAMShift tracker cannot track the markers for n frames, then the marker is labelled as "lost" and all blobs in the mask are searched for the lost marker. Finally, an ellipse is fit to the tracked markers and the center of it is used as the position of the marker.

5.4 Results

5.4.1 Experimental Setup

Preoperatively, all instruments are marked with different colour markers. The colour markers used are biocompatible and sterilization proof. This allows for permanent use of these marked instruments in-vivo during surgeries. Different colour markers were used (red-pink-yellow-blue-green) in three test scenarios as illustrated in figure 5.3. In scenario 1, a single instrument with a single colour

Figure 5.3: *Test scenarios*

marker is tracked in the endoscopic video. In scenario 2, a single instrument with a multi-colour marker is tracked in the endoscopic video. Finally, in scenario 3, multiple instruments marked with different single colours are tracked simultaneously in the endoscopic video. The goal of the experiments is to test the feasibility of the tracker in real-time endoscopic video and the performance of the different individual colours in the three scenarios described above.

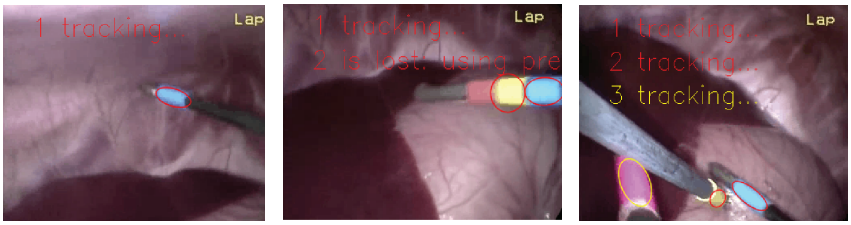


Figure 5.4: *Tracking results*

The tracking system is used in real-time during animal trials. The tracking results are shown in figure 5.4. Different segments of endoscopic video are collected for the different test scenarios. Segments include various environmental distortions, such as, smoke, camera motion and organ occlusions. The method is tested on a total of 10 minutes of endoscopic video material. Ground truth data is generated by manually labelling the instrument appearance in the endoscopic video. For every frame of the video, a 1 is assigned if the instrument appears in the frame and 0 if it does not appear.

5.4.2 Results

The classification results are shown in figure 5.5. The plots show the ROC curves of the detection rates of different colours for the three scenarios. The general results show that colours blue, green and yellow, show high sensitivity and specificity results, both for single and multi-instrument detection, however, pink and red show bad performance as they are completely confused with the abdominal

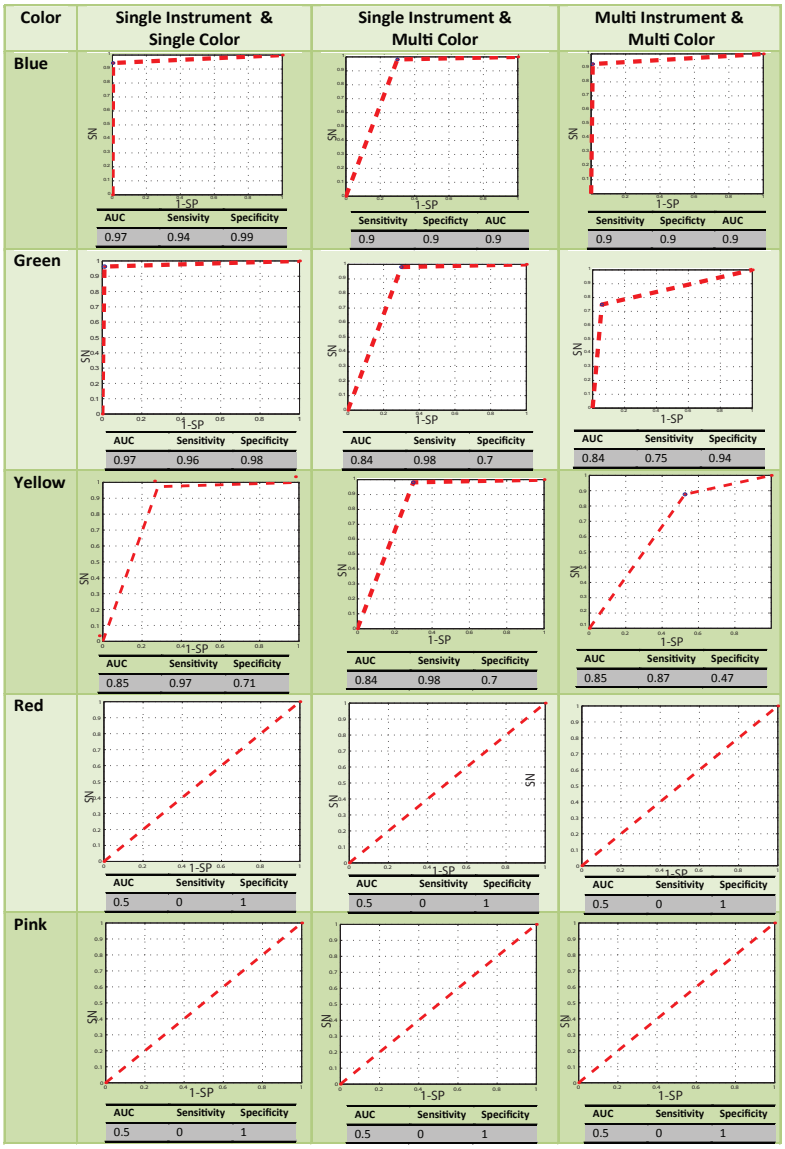


Figure 5.5: *ROC curves.*

background colours. For scenario 1, single colour tracking shows that colours blue and green are the best suited for tracking with the highest sensitivity and specificity results; however, yellow shows less performance as it can be confused with

body fat under specific environmental conditions. Scenario 2 results show that the single instrument coded with multiple colours can be tracked; however the performance is limited to the lowest performance of the colours used, in our case yellow. Scenario 3 results show that simultaneous multi-instrument detection is possible, however, as more colours are used, the confusion increases between those colours, resulting in lower performance of close colours in the hue space such as green and yellow. Furthermore, the results show that colour combinations relatively far in the hue space show better performance than colours relatively close in the hue space. Hence, it is preferred to use a combination of colours with maximum distance in the hue space.

In general, the tracker shows similar performance under various environment distortions as smoke, camera motion and organ occlusion, except for the yellow colour which is sometimes confused with body fat. A typical tracking speed with a single target, tested in a video with 640x480 resolutions, was 55.2 frames/second. The time performance of the system is more than sufficient for a real-time use, and therefore, this algorithm is a strong candidate for a real-time application in-vivo during surgery.

5.5 Conclusions

We presented a real-time, multiple instrument tracker for in-vivo use during surgery. This tracker is shown to cope with varying sizes of instruments, smoke, camera motion and occlusion. Furthermore, the tracker shows a high processing speed, which is sufficient for real-time use.

The marker colours green, yellow, and blue can be used for both single and multiple instrument tracking. For single instrument tracking, the tracker achieves an excellent performance for the colours blue and green. Yellow is, however, occasionally confused with body fat under specific conditions. Multiple colours coding of single instruments is feasible, however, the tracking performance is limited to the lowest performance of the single colours used. Simultaneous multiple instrument tracking is also feasible with the proposed tracker, however, as the colours are closer in the hue space, the confusion increases between those colours. Furthermore, for multiple instrument tracking, colour combinations far in the hue space show better performance than colours close in the hue space. Hence, it is preferred to use colours with maximum distance in the hue space. Red and pink colours are completely confused with abdominal background tissues and cannot be used for tracking in endoscopic video with this tracker.

Future work includes testing on different available biocompatible colours. Furthermore, single and multiple colour coding schemes need to be designed and tested to deal with a high number of different instruments under the strict constraint of high sensitivity and specificity of the tracker. Finally, we aim to test the system feasibility in combination with a workflow information system for a real-time use in-vivo during surgery.

Chapter 6

Intraoperative: detection of surgical outliers

L.Bouarfa and J.Dankelman
submitted article
under the title "*Measuring consensus and detecting outliers from surgical process logs*"

abstract

Purpose: The purpose of this work is twofold: (1) to derive a surgical consensus workflow from multiple surgeries and (2) to detect outliers automatically from a (running) surgery.

Methods: Workflow mining is used in this paper to derive a surgical consensus from multiple surgery logs using tree guided multiple sequence alignment. A process log is directly derived for each surgery from laparoscopic video using an already developed instrument tracking tool. In total 26 surgery logs are used to derive the consensus for laparoscopic cholecystectomy. Finally global pair-wise sequence alignment (Needleman-Wunsch) algorithm is used to detect outliers for running surgeries.

Results: We showed that a generic consensus can be derived from surgical process logs using tree guided multi-alignment. The derived consensus conforms the main steps of laparoscopic cholecystectomy as described in best practices. Using global pair-wise alignment, we showed that outliers can be detected using the consensus and the surgical process log.

Conclusion: We used tree guided alignment to derive surgical consensus and to detect outliers from running surgeries. Detecting outliers in surgery is a valuable tool to analyse the underlying cause and improve surgical practices.

6.1 Introduction

Workflow mining is a technique which aims at improving the workflow modelling process by providing tools for discovering, comparing, and conformance checking of workflow process models [Aal03]. Conformance checking is crucial in the domain of surgery to detect deviations (i.e. outliers) from surgical protocols; detection of outliers after surgery can be used for early error prediction and intra-operative alarming of the surgical team to avoid surgical errors. Moreover, workflow mining is a tool to enrich and extend surgical protocols with frequently occurring practices. Hence, workflow mining is a valuable tool for modelling surgical workflow.

Workflow variability is inherent to the domain of surgery [GM10]; it is caused by uncertainty in patient anatomy, unexpected complications, surgeon cognition and situational awareness. Moreover, unlike organizational workflow, surgical workflow lacks the direct human computer interaction, which makes automatic activity logging a necessary step before modelling [Bou11b]. Therefore, workflow activity monitoring in the OR is non-trivial. Implementing surgical workflow analysis requires an approach for dealing with variability and monitoring of surgical activities.

In laparoscopic surgery, the surgeon inserts long and thin instruments in the abdominal cavity through small incisions while an assistant holds the laparoscopic camera. Because of the small incisions this is considered to be less traumatizing for the patient than open surgery. The laparoscopic video is the main input and feedback signal of the surgeon to execute the surgical task.

In our previous work Laparoscopic Cholecystectomy (LapChol) procedures are considered for probabilistic workflow modelling [Bou11b]. LapChol is a highly standardized surgical procedure in which a patient's gallbladder is removed in case of inflammations. The presented approach in [Bou11b] deals with the variability of LapChol workflow using a probabilistic Markov-based approach to detect the different surgical workflow steps. The high standardization of the LapChol procedure, together with the achieved results with the probabilistic approach, challenge us to test a framework for generating a consensus and detecting outliers for LapChol surgeries using only workflow mining without prior knowledge of the workflow itself.

This paper presents a workflow mining framework for offline utilization in the OR intended for quantitative mining of laparoscopic surgical workflows. This workflow mining framework is shown to allow for the variability of the surgical procedures. The framework can be applied to any laparoscopic procedure, but is currently only evaluated for laparoscopic cholecystectomy.

The general framework is presented in section 6.2. Section 6.3 shows how to generate a process log from laparoscopic video. Section6.4 presents the workflow mining algorithm for generating a consensus workflow and Section6.5 shows how outliers can be detected using the consensus derived in section 6.3. Finally, section 6.6 concludes this paper and highlights our future work directions.

6.2 General framework

The first goal is to generate a consensus workflow from multiple surgery logs. Secondly, we aim to detect outliers from each surgery by comparing its activity log to the general consensus workflow.

Figure 6.1 illustrates the framework of the workflow mining system for LapChol, in the following steps:

1. The first step is to generate process logs form laparoscopic video using the real-time tracking system already developed in our previous work [Bou11a]. The tracking system generates a process log as described in section 6.3 with an entry for each time an instrument is used.

2. The second step in our framework is to make use of many process logs of the same type of surgery to derive a surgical consensus. This step is performed offline and uses many surgery logs. To generate the consensus a multi-alignment algorithm is used as described in section 6.4

3. The final step is to detect outliers, also known as anomalies, during surgery. This step takes place at the end of the surgery, when the process log as it is generated by the tracking system, is compared to the consensus. This is described in section 6.5.

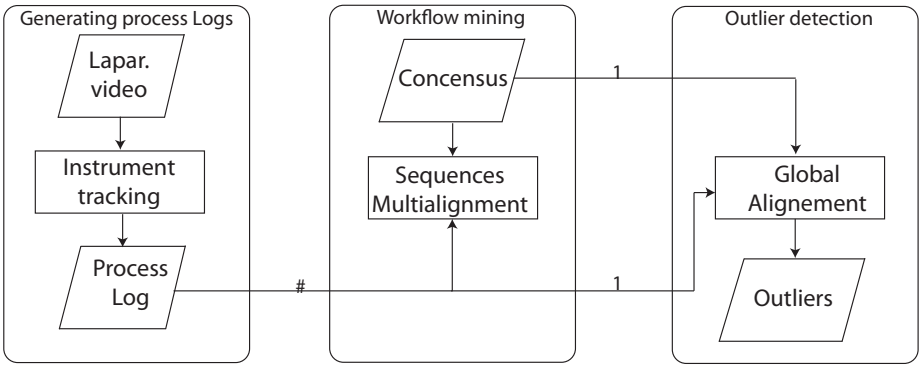


Figure 6.1: *General framework for workflow mining and outlier detection in surgery*

6.3 Generating process log from laparoscopic video

Laparoscopic video is used in this section to generate the surgery log. Therefore a tool is designed to detect and track single and multiple instruments in laparoscopic video using biocompatible colour makers. Figure 6.2 illustrated the

tracking algorithm. Initially, the colour markers placed on the instruments are segmented from the background using colour information. Afterwards, the markers are tracked in the segmented regions to extract their trajectories and produce their process log.

For the segmentation step, the colour markers are selected by using a HSV (hue, saturation and value) colour map such that they are positioned with maximum distance relative to each other in the hue space. To build a marker-colour histogram in hue-saturation colour space different images are captured under various illumination conditions and from different parts of the laparoscopic video. Finally, a binary mask, is created by thresholding the pixel probabilities. Morphological closing (dilation) is performed on the resulting mask to fill possible gaps in the masks and to extend the marker search region for robustness.

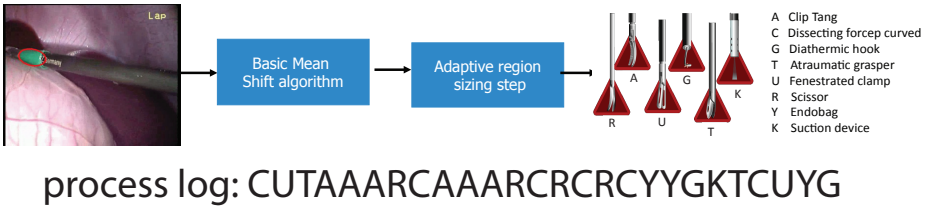


Figure 6.2: *Instrument tracking tool to produce a surgical process log from laparoscopic video*

For the tracking step, first step blob is performed on the mask, built in the previous marker segmentation step to group foreground pixels. Afterwards, the markers that are positioned inside the blobs are tracked using the OpenCV implementation of a Continuously Adaptive Mean Shift (CAMShift) algorithm [ope]. A separate tracker for each marker is used together with its colour model. Finally, a Kalman filter [Wel95] is employed together with the CAMShift tracker in order to cope with situations when the markers are not visible, lost due to the image noise or occluded by other instruments or abdominal tissues. At this final step, an ellipse is fit to the tracked markers and the centre of it is used as the position of the marker.

The output of the tracking tool is a process log file with the symbol of the instrument used at each entry and the duration of its use as illustrated in figure 6.3. In this paper we discard time information from the analysis as it can always be retrieved afterwards, and focus only on the process log with the instrument symbols presented in the first column of fig 3.

Instrument symbol	Start time	End time
C	08:09:27:20	08:10:15:40
U	08:11:07:12	08:11:07:12
T	.	.
A	.	.
A	.	.
A	.	.
R	.	.
C	.	.
A	.	.
A	.	.
A	.	.
R	.	.
C	.	.
R	.	.
C	.	.
R	.	.
C	.	.
Y	.	.
Y	.	.
G	.	.
K	.	.

Figure 6.3: *An example of a process log generated by the instrument tracking tool from [Bou11a]*

6.4 Workflow mining : Generating surgical consensus using multi-alignment of individual process logs

A consensus workflow for a specific procedure can be constructed using expert opinions, however this can require a lengthy debate, with no guarantee to reach a final consensus. Another option is to automatically derive the consensus workflow from multiple individual process logs. This has as advantage that the consensus will most reflect the reality and that no expert opinion is required. In this section an approach to the automatic derivation of consensus from individual process logs is presented.

Sequence alignment is used in bioinformatics to find overlapping or similar sequences of DNA, RNA, or proteins and to identify important relationships. It deals with the problem of grouping together sets of sequences to identify regions of similarity between the sequences. Gaps are inserted between the elements of the sequence to align similar characters in successive columns. Taking inspiration from biological sequence alignment, Jagadeesh et al. [JCB10] proposed to apply this technique for process mining. To derive the consensus we performed progressive multiple alignment for 26 surgical process logs guided by a scoring tree.

Figure 6.4 illustrates the multiple sequence alignment used to derive the surgical consensus form 26 surgical process logs. To generate this alignment, first, a distance-matrix is constructed by calculating pair-wise distances between all process logs. This matrix is used to build a guide tree using a neighbour joining algorithm. In the next step, the sequences are progressively aligned using the guide tree that defines the order in which the sequences are aligned in the align-

ment step. Finally, The consensus values are calculated from the aligned sequence weighted using scoring matrix whose values are the average Euclidean distance between the scored symbol and the M dimensional consensus value where M is the size of the alphabet.

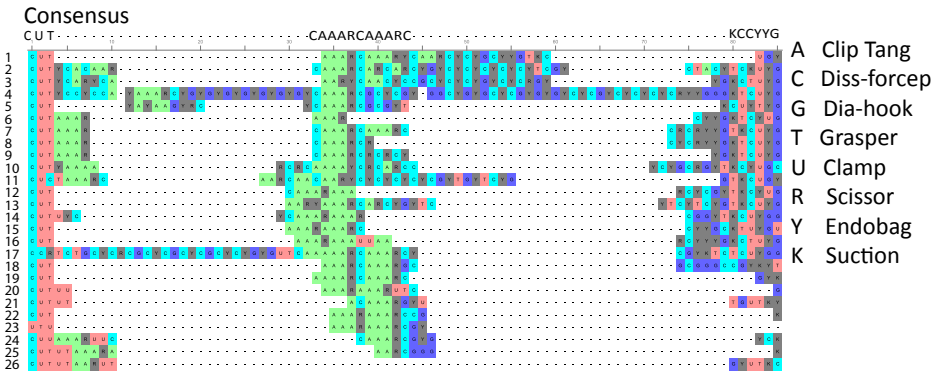


Figure 6.4: *Using tree guided multi-alignment to generate consensus from surgical process logs*

The derived consensus shows the main steps of LapChol workflow. The primary step of the surgical procedure is to dissect the area which includes the bile duct and the cystic artery (Calot's triangle). The main instruments used in this step are represented by the characters C and T. When both structures are clearly visible, each of them is clipped with three clips, (AAA), followed by dissecting and cutting both structures between the clips with laparoscopic scissors (RC). The following step is dissection of the gallbladder using the dissection device(C). During and after the dissection of the gallbladder, the surrounding abdomen area is cleaned using the suction device (K). To remove the dissected gallbladder, a salvage bag is inserted into the abdomen, the gallbladder is packed up into the bag and the bag extracted together with trocar (represented by Y). Finally, the surgical area is explored to detect and stop any bleedings. A drainage is inserted through a trocar hole and all instruments are removed.

6.5 Outlier detection using global alignment

Given an arbitrary Lapchol process log, the deviation from the general consensus is computed. The differences between the two sequences are calculated by a global pair-wise sequence alignment (Needleman-Wunsch) algorithm.

As illustrated in figure 6.5, the algorithm inserts gaps into the process log in order to align it to the consensus. Those gaps are deviations (i.e. outliers) of the surgery from the consensus. Note that all surgeries from the previous figure 6.4 contain outliers (i.e. are different from the consensus), which is to

be expected given the variable nature of surgical workflow. Those outliers are commonly simple variations in the execution of the surgical procedure, but can also represent serious complications or errors. Figure 6.5 provides a description of the outliers detected which helps to describe the surgery in more detail.

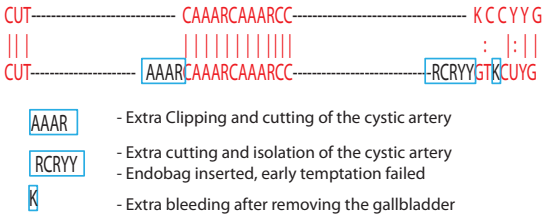


Figure 6.5: *Using global alignment to detect outliers*

6.6 Conclusion

In this paper, a new approach for deriving surgical consensus from real-life surgeries is proposed. Surgical logs can be derived real-time from surgeries using our previously published tracking software [Bou11a]. We have showed how guided multiple sequence alignment can be used to derive a workflow consensus that can be used as the reference workflow for detecting outliers during surgery. The derived consensus contains the major steps of the Lapchol surgery. We also showed how outliers can be derived using pair wise global alignment. Although it is important to monitor the compliance of surgical workflow with the consensus, it is also interesting to look at the outliers in the general workflow and the underlying reasons, as they can represent patients with beneficial modifications in treatment. Finally, time information about the outliers can be obtained from the original process log.

The calculation of the consensus is sensitive to adding new surgical logs, for future work, we recommend to use more data and to use more advanced sequences clustering algorithms to deal with the variations in the logs. We also recommend for future work the online computation of outliers during surgery and the classification of the detected outliers using prior knowledge structured in a decision tree to allow for automated annotation of outliers in surgeries.

Chapter 7

Postoperative: prediction of recovery time

L.Bouarfa, D.M.J.Tax, J.M.Ehrenfeld, B.Rothman, J.Dankelman
Published in the Journal of Artificial Intelligence in Medicine, (2011)
under the title "*Length of Stay in the Post Anaesthesia Care Unit - Can it be estimated?*"

abstract

Objective: The post anaesthesia care unit (PACU) is a costly but important peri-operative healthcare necessity. Accurately predicting patients' length of stay (LOS) in the PACU may lead to cost savings and a number of operational benefits.

Methods and material: After receiving the institutional review board (IRB) approval, electronic data from a 10-year period were collected from the peri-operative data warehouse at the Vanderbilt University Medical Centre. Data included case demographics, intra-operative parameters, medications, patient co-morbidities, and surgical factors. Cases with missing data were removed from the analysis. A linear regression method was employed along with ordinary least square regression and 'least absolute shrinkage and selection operator' (LASSO-) regression that allowed data discretization. A forward feature selection approach was then used to identify and rank factors impacting PACU LOS.

Results: After pre-processing we used data from 53,464 patient encounters. The least square regression with forward feature selection provided better performance than the LASSO technique, requiring only ten features to provide a maximum average improvement of 12 minutes compared to the *mean baseline* [1]. This result is achieved without including the surgeons and anaesthetists in the regression model. A 6-minute performance gain occurred when surgeons and anaesthetists were added to the model, resulting in a total improvement of 18 minutes from the mean baseline by using all features.

Conclusions: PACU LOS can be predicted by peri-operative factors with an improvement of 12-18 minutes compared to using the mean baseline. If this prediction is updated with online information, mainly by monitoring post-operative oxygen saturation, future work could lead to real-time LOS algorithms based on peri-operative factors to predict, manage and possibly intercept anticipated, prolonged PACU LOS.

[1] The mean baseline is the mean absolute deviation $\epsilon_b = \frac{1}{N}\sum_{j=1}^{N}(t(j) - \frac{1}{N}\sum_{j=1}(t))$ from the total mean. It represents the average of all the distances between the measured PACU LOS for each patient and the the average LOS of all patients

7.1 Introduction

The post anaesthesia care unit (PACU) is a costly healthcare necessity. Patients of varying illness severity progress through anaesthetic and surgical recovery phases at different paces after surgeries of varying complexity. Appropriate post-operative care requires high staff-to-patient ratios, and staff salaries represent a significant percentage of total cost to operate a PACU [Tes99].

Recovery at the PACU is an on-going process that begins when the intraoperative period has ended and continues until the patient returns to their preoperative physiological state. This process is divided into three phases. The early recovery (Phase I) is the transition period from a totally anesthetized state to the recovery state of the protective reflexes and motor functions. The intermediate recovery (PhaseII) is the period during which the patient is prepared for self-care by family members as intensive nursing care is no longer needed. Finally, in the late recovery (Phase III) patients who require extended observation are monitored for extra time (e.g. extra overnight, home monitoring) until the patient is back to his preoperative functional status [Twe08].

PACU LOS (post anaesthesia care unit length of stay) is commonly defined as "the number of minutes between the time the patient arrived in the PACU and the time the patient departed the PACU" [Dex95a]. This time may include Phase 1, 2 and 3 of the recovery process. When PACU LOS for each patient increases, the number of patients in the PACU increases, potentially requiring greater numbers of staff. This can have a dramatic effect on healthcare delivery costs [Tes99]. Cost comparisons for 2 hours in a PACU are equivalent to a 24-hour stay in a hospital ward [Dex95a; Wad98].

Macario et al. recognized the need for cost analysis as early as 1995 [Mac95]. Their work identified PACU costs, that are proportional to volume and related specifically to patient care and variable direct costs, to be 32.7% +/- 0.1% of a hospital stay. Dexter et al. showed that nearly all of this cost was labour expense and that admission distribution could limit a hospital's ability to reduce costs in this area [Dex95b]. A time-study demonstrating that 72.5% of PACU nursing time relates to direct patient care reveals the potential magnitude of this labour [Coh99]. Translating a decrease in PACU LOS into definitive cost savings for a facility has been a challenge. Dexter et al. has demonstrated that time savings in the PACU is not enough. The pay structure for nursing staff [Dex99] and the ability to decrease the number of nurses required to care for a peak number of PACU patients [Dex95b] appears to be significant. The latter factor can be affected by optimal case sequencing [Mar06; Dex05], efficient PACU discharges [Dex95b], and by-passing Phase I recovery [Dex99]. This decreases the necessary nurse to patient ratio from 1:2 to 1:3. Without considering all of these factors, the expense required to decrease LOS does not generate the expected savings. The ability to prospectively discriminate between a customary and prolonged patient length of stay (LOS) in the PACU provides data that could allow peri-operative managers to anticipate and allocate resources accordingly. In the event that

resource constraints become a limiting factor, this information could be escalated to other peri-operative managers to identify potential bottlenecks and attempt to improve the flow of patients through the peri-operative process. This data could be provided to front line managers through a notification system or through electronic case boards.

Prolonged LOS studies have been previously performed at a variety of large, small, and ambulatory facilities. Study types range from simulations based on historical data, retrospective and prospective depending on the study hypotheses. The previously identified causes for prolonged LOS are multifactorial. Patient-centred reasons may include anaesthetic technique [Mur04] and type and length of surgical procedure [Chu99], patient morbidities, and post-anaesthetic and post-surgical factors [Mar03]. Post-operative adverse events also appear to prolong LOS [Coh99; Chu99; Sam06]. System-centred reasons relate to delays in coordinating, synchronizing and mobilizing the multitude of resources necessary to qualify and physically move the patient to the next stage of care [Tes99; Wad98; Mar03; Sam06; Bro08]. A review of the literature revealing definitions for being appropriate for discharge and actual discharge have been used to identify systemic-centred delays [Wad98; Mar03]. For the purpose of delays entering the PACU, the LOS has also been defined as the number of minutes from the time the PACU bed was requested until the time of PACU discharge [Dex95a; Sch09].

The Vanderbilt University Medical Centre has collected recovery data for the last 10 years. This study aims (1) to assess different regression models for predicting the PACU LOS from pre- , intra- and post-operative patient data and identify the best method for prediction, and (2) to identify and rank the significant parameters of a prolonged PACU LOS from a large sample of patients possessing a variety of comorbidities and scheduled for a variety of surgical procedures.

7.2 Materials and methods

This study was conducted using data collected by the Vanderbilt's peri-operative information management system $VPIMS_{TM}$, a point-of-care database that supports the care documentation, financial, and quality improvement processes of the peri-operative enterprise.

7.2.1 Data

This data was collected from the Vanderbilt University Medical Centre, an 800+ bed tertiary-care hospital, at which approximately 65,000 anaesthetics are now performed each year. PACU times for this study were limited to two main adult PACU suites that are currently 11 and 35 rooms. PACU slots sum up to 46 beds. Per year, approximately 24,300 patients are admitted to Phase 1, and the remainder are either admitted to Phase 2, Phase 3, or directly to the ICU.

Using Vanderbilt's peri-operative information management system, $VPIMS_{TM}$, we collected PACU admission times, discharge ready times (appropriate for discharge), and actual discharge times for the two PACUs between January 1, 2000 and September 30, 2010.

Pre-operative, intra-operative, and post-operative features included in this study, described in Appendix B, were selected because of their significance or consideration in prior studies, or they were deemed to be potentially relevant with respect to the nature of the patient, anaesthetic, or surgery.

Pre-operative features include patient demographics (age, gender, and ASA physical status classification), medical history (e.g. smoking history, diabetes, cardiac failure, substance abuse) and anaesthetic data including providers, planned type of anaesthetic, and elective vs. emergency surgery. Intra-operative features include medications administered to the patient during surgery, room staff level of training, invasive monitor use and blood administration. Post-operative features include duration of surgery and anaesthesia. In the analysis, the features have been considered as ordinal or binary variables. Binary features include most of the features in the dataset. Ordinal features include age, Body Mass Index (BMI), operative time, and duration of anaesthesia. The feature details are described in the supplementary material of Appendix B.

The PACU LOS is measured as the time t from patient's admission to the PACU until the patient's discharge from the PACU (i.e. including Phase I to III). In this dataset the PACU LOS has an average of 190 minutes and a standard deviation of 206 minutes. Within this period, a patient may be in Phase I, Phase II or Phase III of discharge, representing their recovery progression and respective nurse to patient care ratios required throughout their recovery process.

7.2.2 Statistical analysis

Statistical analysis to predict the response variable t, representing the PACU LOS, is divided into three stages. 1) The pre-processing stage filtered the data and mapped it to another scale (i.e. logarithmic) suitable for analysis. 2) The regression stage assessed regression models to predict optimally the PACU LOS from the available features. 3) The feature selection stage applied the regression to identify the predictive features of PACU LOS. Both regression and feature selection are performed using the statistical toolbox PRTools for matlab [Dui07a].

Preprocessing

Scatter plots of the data revealed the presence of significant biases. The first bias is the patients in the PACU who remained as 'overnight stays', generally due to the unavailability of floor beds. These prolonged stays range from overnight to two days. It is common practice at Vanderbilt to continue documenting their stay in the PACU in the peri-operative electronic chart as a Phase III patient. The second bias was a 'short stay' bias due to the performance of outpatient surgeries

requiring little more than a single set of vital signs to meet PACU discharge criteria. Finally, many cases with missing values were present. To allow sound regression analysis, the pre-processing excluded all these cases: Of the 311,374 patients included in the study, approximately 83% were excluded, leaving 53,464 patients in the final dataset.

The response variable t was then normalized. The distribution of the response variable t in this dataset follows a lognormal distribution, shown in the Figure 7.1 left subplot. To fit the regression function, it is most convenient to use a symmetric error, rescaling the response variable to a logarithmic scale. This avoids the influence of large response values in the right hand tail of the distribution and a high bias predicting the average PACU time. After logarithmic normalization, the response variable distribution in the Figure 7.1 right subplot was obtained. The distribution is close to normal, making it very suitable for fitting a (linear) regressor using a symmetric error.

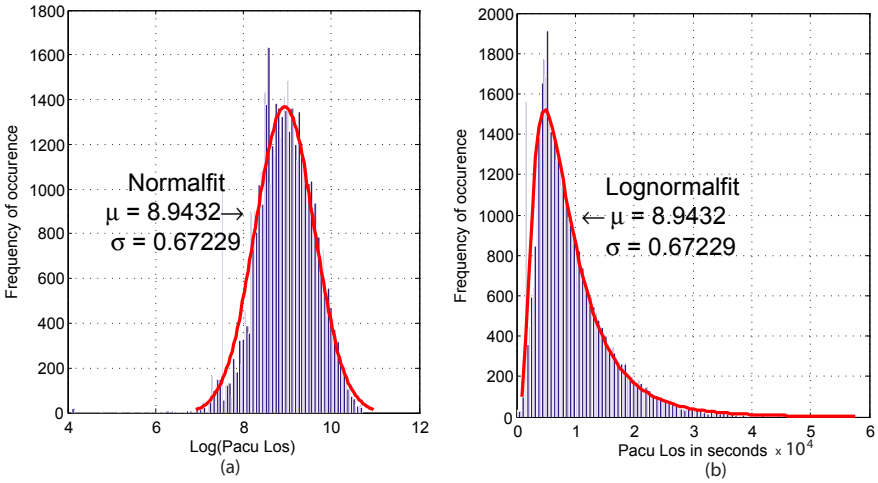


Figure 7.1: *(a) Distribution PACU LOS; (b) Distribution log (PACU LOS)*

Feature selection

Feature selection was used to reduce the large feature set to a small subset allowing for optimal regression performance and less noise sensitivity. The forward feature selection was chosen because others become extremely time consuming with larger feature sizes. Forward feature selection extends a preliminary subset of features to that feature for which the performance improves most. Here, the mean square error (MSE) criterion function was used to assess this subset of features. The

MSE is defined by:

$$\epsilon_{max} = \frac{1}{N}\sum_{j=1}^{N}(log(t(j)) - log(\hat{t}(j))^2 \tag{7.1}$$

The mean square error measures the average of the squares of the error between the measured PACU LOS $t(j)$ and its estimated value $\hat{t}(j)$.

Regression

Linear regression predicted the value of the PACU LOS t based on a linear combination of feature values. The regression model used is: $\hat{t} = w_0 + w_1x_1 + w_2x_2 + ... + w_px_p$, where $\hat{t}$ is the estimated PACU LOS, $\vec{w} = (w_1, w_2, ..., w_p)$ are feature values and $x_{1..p}$ are the weights. Weighting of features can be used to remove irrelevant or redundant features and rank the remaining features.

Nonlinear regression is an alternative to linear regression. In general, nonlinear regression is more flexible with more freedom to adapt to the data, but it also requires more data to fit the parameters. As noisy features are abundant and noisy, overfitting and reduced performance becomes problematic. Furthermore, choosing the functional form of the nonlinearity is often not straightforward, and there was no clear structure visible in this noisy data. Applying a nonlinear kernel smoothing approach results in a higher error (mean absolute deviation) than the linear model. Finally, while nonlinear regression assigns higher order weights to features, they cannot be used to assess the importance of the different features.

Two linear regression techniques have been utilized in this study for estimating PACU LOS t ; the ordinary least square regression and the least absolute shrinkage and selection operator (LASSO) [Tib96]. Both approaches minimize a different error on a given training set, consisting of N pairs of features and targets $(\vec{x}_i, t_i), i = 1, ..., N$. In the ordinary least squares regression, weights are found by minimizing the squared error between the prediction and the true value of the response variable.

When the number of training pairs is small, or the number of features is high, the least squares solution can overfit. A regularized least square is used to avoid overfitting. Adding an extra term $\lambda|\vec{w}|^2$ to the squared error suppresses solutions in which some weights are very large, reducing the probability of a few noisy features will completely determine the solution. Alternatively, the LASSO tries to avoid overfitting by introducing an additional constraint that $|w_1| + |w_2| + ... + |w_p| \leqslant \Lambda$, so weights are not punished with a squared error, but with an absolute error. Larger weights are punished less, relatively, and importantly many weights are optimized to be exactly zero. LASSO performs an implicit feature selection since features with a zero weight are not needed in the regression solution. The advantage of this approach is that features are selected purely based on the utility in the regressor itself, and no external or suboptimal criterion needs optimisation as an intermediate step. The least square regression, however, was performed

with an external forward feature selection as described in section 7.2.2. Other regression techniques involve discretizing the time range of the PACU LOS and predicting each level separately using classification techniques. However, our goal is to predict the PACU LOS as a real value and not as a discrete level.

To evaluate the performance of the least square and LASSO regression, the mean absolute deviation ϵ_r was used to measure the amount of deviation (variation) in minutes of the predicted $\hat{t}(j)$ from the measured PACU LOS $t(j)$:

$$\epsilon_r = \frac{1}{N}\sum_{j=1}^{N}(log(t(j)) - log(\hat{t}(j)) \tag{7.2}$$

Where $t(j)$ is the measured PACU LOS, $\hat{t}(j)$ the predicted PACU LOS and N is the number of instances in the dataset. The mean absolute deviation allows for expressing the error in minutes between the measured and the predicted PACU LOS.

7.3 Results

We conducted both regression and feature selection as explained above using the statistical toolbox PRTools [Dui07a]. Using forward feature selection we measured the selection process stability when different test sets drawn from the original dataset were applied. We then identified the subset of features that allow the prediction of t with accuracy comparable with the full feature set. Using feature curve analysis we evaluated the performance of the regressors in fitting the PACU LOS data. The optimal regression technique predicting the response variable t and the number of samples required for optimal training is described below.

7.3.1 Feature selection results

Forward feature selection was used to identify and rank the relevant features for predicting PACU LOS. The feature stability was measured by repeating the selection experiment 100 times, randomly selecting half of the data for fitting (i.e. training), and the other half of the data for testing. The first experiment excluded surgeons and anaesthetists from the selection process. Figure 7.2 illustrates the top ten ranking levels using forward selection. A feature subset is considered stable when each ranking level is dominated by few features. Forward selection of the top ten ranking levels proved stable for the subset of features including general anaesthesia, CRNA, neostigmine, regional anaesthesia, anaesthesia time, age, cardiac failure, SRNA, arterial line use, gender and operative time.

The second experiment included surgeons and anaesthetists in the selection process as illustrated in Figure 7.3. The top ten ranking levels also include two anaesthesiologist and five surgeons. Surgeons and anaesthetists are important

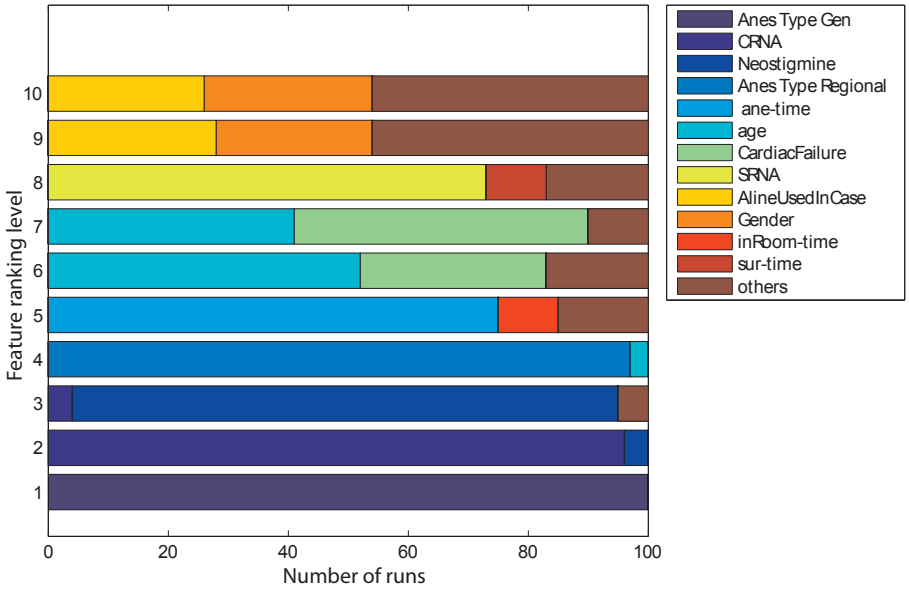


Figure 7.2: *Feature stability for the top ten feature subsets using forward feature selection excluding surgeons and anaesthetists*

features strongly affecting the feature selection mechanism in its top ten ranking. This point to the potential that practice and procedure variation can vary enough to impact PACU LOS. The other main observation is that feature selection becomes unstable from level 4 onwards when adding surgeons and anaesthetists. This result suggests that, although surgeons' and anaesthetists' features may positively contribute to predicting PACU LOS, they introduce high variations in the dataset making the feature selection mechanism unstable to identify the relevant features for predicting PACU LOS.

The forward feature selection process included ASA classification and emergency case status. The absence of the classification from the top ten is of particular interest and is consistent with bodies of work that find little correlation between pre-operative patient status and PACU LOS [Chu99].

7.3.2 Regression results

Feature curve analysis was used to examine the relationship between the mean absolute deviation of the regressor and the size of the feature set. Least-square and LASSO regression methods were the best two linear regression techniques in predicting t.

Feature curve analysis plots the mean absolute deviation against the number of

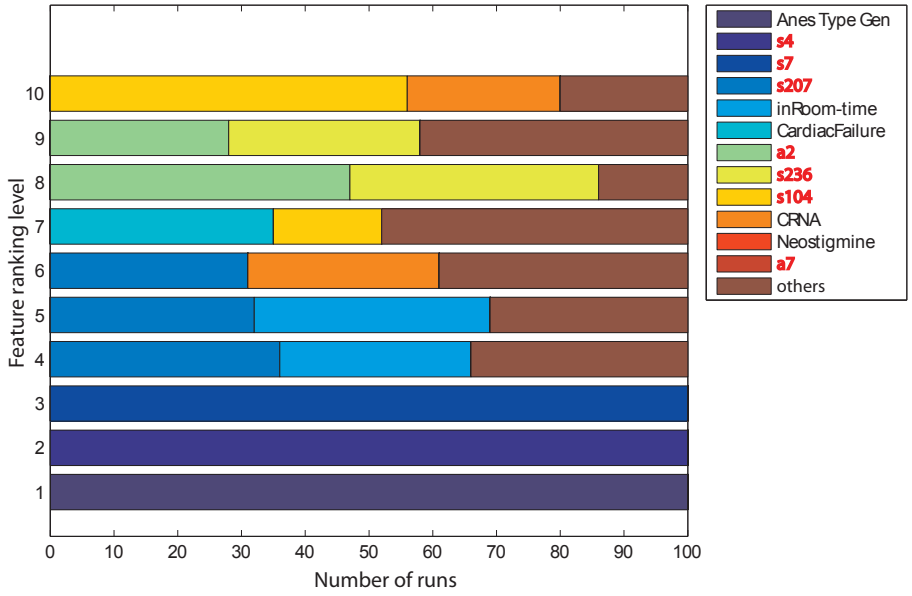


Figure 7.3: *Feature stability for the top ten features using forward feature selection including surgeons and anaesthetists, note that the added features are given a red label*

features p used in the training process. A steeply decreasing feature curve suggests that better performance can be obtained when more features are available. A flattened feature curve suggests that the regressor is already well trained, and more features would not significantly improve the performance of the regressor. Feature curves typically report which regression technique is suitable for small feature set sizes, which regression technique has the most promising performance and how large the feature set size should be for an optimal performance of the regression technique. In this experiment, a fixed training size is used, half of the data is used for training the regressor, and the other half for testing.

Figure 7.4 represents the feature curves of the least square and LASSO regressions, excluding features representing surgeons and anaesthetists from the data. The total feature space includes 60 features. The red dotted line represents the mean baseline, which is the absolute error between the measured PACU LOS (t) of each patient from the dataset and the total average PACU LOS from the dataset. The regression curves represent the varying absolute error against an increasing feature size from 1 to 60. Both the least square and the LASSO show a minimal prediction error of 68 minutes, but the least square regression performs better than the LASSO flattening out after 10 features, where the LASSO needs at least 20 features to achieve the same minimal error. Hence, a maximum average improvement of 12 minutes from the baseline can be achieved by using the features

without considering the surgeons and anaesthetists in the regression models. To

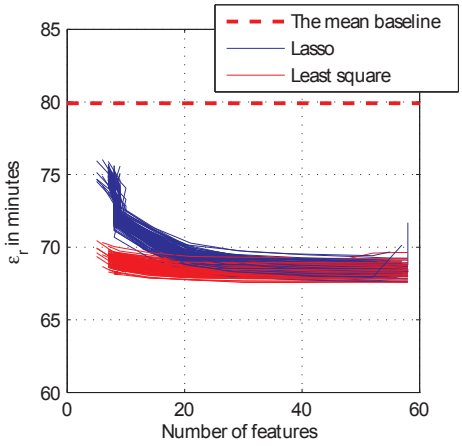


Figure 7.4: *Feature learning curves for Linear and Lasso regression, number of runs=* 100

investigate the existence of performance gain when surgeons and anaesthetists are added to the regression model, the feature curves in Figure 7.5 represent the least square and the LASSO regressions when all features are used. It demonstrates a further performance gain of 6 minutes when all surgeons and anaesthetists are added to the model. Note that the full feature space is comprised of 425 features, of which 365 represent surgeons and anaesthetists responsible for the surgery, and flattening of the feature curves occurs at approximately 200 features. A total improvement of 18 minutes can be achieved from the baseline by using all features.

7.4 Conclusion & Discussion

The study makes several contributions, most notably, the idea of predicting LOS from readily available pre-, intra- and post-operative patient data to increase PACU efficiency. To our knowledge, this is the first attempt to predict PACU LOS as a real value in the literature.

After removing biased cases and those missing data, the least square and the LASSO linear regression methods revealed ten features that are easily acquired from an Anaesthesia Information Management System that can be used to sufficiently estimate the number of minutes a patient's will spend in PACU.

Feature curve analysis determined that the least square with forward feature selection provided better performance than LASSO requiring only ten features for optimal prediction performance. Feature curve analysis also indicated that these

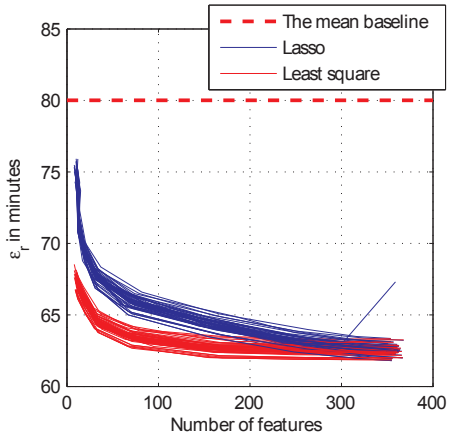


Figure 7.5: *: Feature learning curves for Linear and Lasso regression including surgeons' and anaesthetist's features, number of runs*= 100

ordinary regression methods are best suited for this noisy dataset, outperforming nonlinear regression. A maximum average improvement of 12 minutes from the baseline can be achieved by using the features without considering the surgeons and anaesthetists in these regression models. A subset of only 10 features was required from the complete set to predict the PACU LOS. A 6-minute performance gain occurred when surgeons and anaesthetists were added to the model, resulting in a total improvement of 18 minutes from the baseline by using all features.

Some features such as prolonged operative and anaesthesia times and receiving general anaesthesia have been established in the literature as procedure related indicators that predict PACU LOS. Likewise, patient-centric factors - age, past smoker, and cardiac failure - identified here have also been identified as predictors of prolonged PACU LOS. Neostigmine administration relates to the use of muscle relaxants whose use has been shown to prolong LOS. Arterial line use likely relates either to the complexity of the case or the patient's pre-operative conditions, or both. The significance of gender and CRNAs/SRNAs are unknown and will require further investigation.

Interestingly, specific surgeons and anaesthesiologists were also predictors of a prolonged LOS. While these features made the feature selection mechanism unstable, the identification of two anaesthesiologists and five surgeons in the top ten ranking levels point to the potential that practice and procedure variation can vary enough to impact PACU LOS.

The ability to predict patient's PACU LOS from peri-operative factors could be invaluable in prospective and real-time bed management in the PACU, OR, and hospital wards. If this prediction is updated with online information, mainly by monitoring post-operative oxygen saturation, future work could lead to real-time

LOS algorithms based on peri-operative factors to predict, manage and possibly intercept predicted, prolonged LOS.

This paper makes use of straightforward linear regression methods as the data contains many features, has no clear structure, and has a high amount of variability. Nonlinear models (e.g. kernel smoother) offer many advantages to implicitly detect complex nonlinear relationships between the variables. However, they are prone to over-fitting and can offer less insight into the rationale of the results. Future work may include testing the data using non-linear approaches to test if the performance can be further improved.

Chapter 8

Discussion and Conclusions

8.1 Summary and discussion of results

This thesis has proposed and demonstrated the application of pattern recognition tools to log, assess and predict surgical workflow parameters. The thesis did not directly contribute to reduce errors and safety in the OR. However the tools developed in the thesis can be used to support standarization of surgical workflow to both reduce errors and support surgical planning. We have used different types of surgical data to predict outcomes about safety and efficiency of surgeries. In this section the combined main results and conclusion for the Chapters 2-7 are given.

Evidence-based justification of surgical practice is becoming increasingly relevant to avoid adverse events. Unlike other medical treatments, surgery is a skill-dependent, multistep procedure. This makes evidence-based studies in a traditional RCT framework difficult to be designed. We proposed pattern recognition as an alternative approach in Chapter 2 which allows us to estimate the safety, effectiveness and also efficiency of a surgical treatment for individual patients, using the available biased, noisy and incomplete data. Although it does not provide the same level of evidence as an RCT, it allows for variations in surgical practice and patient anatomy.

During pre-operative planning, demographics and comorbidities are collected daily from patient's pre-operative history and physical examinations. In Chapter 3 we have evaluated different classifiers based on the available evidence from pre-operative data to distinguish between patients with complex and simple surgeries. Experiments showed that intraoperative complexity can be predicted before surgery from readily available preoperative data with an accuracy up to 83% using an LDC or SVM classifier. This study also showed that patients, with inflammation, wall thickening, male sex and high BMI, tend to be at high risk for complex laparoscopic cholecystectomy surgeries. The trained classifiers can be used in pre-operative setting to optimize the preoperative planning by taking the necessary

precautions for the predicted complex surgeries.

By measuring surgical workflow and matching it with well-defined workflows (i.e. best practices), meaningful connections between structure, function and mechanism can be made. In Chapter 4, we have used instrument signals to detect surgical high-level workflow activities during surgery. Experiments on a noiseless dataset of ten surgeries show that it is possible to recognize surgical high-level tasks with detection accuracies up to 90% using instrument signals. By detecting the used surgical instrument on the laparoscopic feed, the proposed framework can detect the phase of the surgery. The presented framework can be used in an intra-operative setting to log the surgical activities for workflow assessment purposes, and to optimize the peri-operative planning by real-time updates on the expected finishing time of surgeries.

To support the design of workflows consensus from real practice, the starting point for workflow mining is to generate the so-called "workflow log" containing measurements of surgical activities as they have been executed. In Chapter 5, we have developed a tracking system to detect and track instruments in endoscopic video using biocompatible colour markers. The system can track one or multiple instruments in the endoscopic video and generate a workflow log of the surgery. Experimental results show that the proposed method can be run in real-time. Moreover, it is robust to partial occlusion and smoke. The system shows high sensitivity and specificity results for blue, green and yellow colours. The tracking system in combination with the workflow segmentation system can be used in a fully automatic way for real-time activity logging during surgery.

Variations in surgical workflow are considered a medical necessity. Acknowledging these variations, leaves us with a valid concern on how adequate the treatment of patients at a certain point of care is [GM10]. The lack of consensus about how a surgical problem should be addressed leaves surgical practice dependent on the surgeon's individual experience and skills. In Chapter 6 we have presented a new approach for deriving surgical consensus from running surgeries. We have shown how sequence multi-alignment can be used to derive a generic consensus that can be used as a ground truth for detecting outliers during surgery. The derived consensus is shown to conform to the main steps of the laparoscopic cholecystectomy surgery as defined in surgical best practices. Although it is important to monitor the compliance of surgical workflow with guidelines and standards in itself, it is also interesting to measure the (frequency of) deviations (e.g. outliers) from the general consensus. For these deviations, the underlying reasons can be analysed by focusing on "process deviations" (concerning modifications in treatment). We also showed in Chapter 6 how outliers can be derived using pairwise global alignment. This approach in combination with the tracking system can be used to automatically monitor the compliance of surgical workflow with the consensus.

Finally, we have applied regression analysis on peri-operative data to predict the length of stay (LOS) of patients in the Post Anaesthesia Care Unit (PACU) in Chapter 7. PACU LOS can be predicted by perioperative factors with an im-

provement of 12-18 minutes compared to using the mean baseline. The regression model can be used in a peri-operative setting to predict the PACU LOS right after the surgery. The predicted values can be used to optimize the planning of time and resources needed for the individual patients at the PACU unit. If combined with real-time postoperative oxygen saturation monitoring during the recovery at the PACU unit, future work could lead to real-time update of the PACU LOS prediction, resulting in a tool to manage prolonged PACU LOS.

8.2 Future research directions

This thesis sets a first step towards the processing of surgical data using advanced pattern recognition tools to improve safety and efficiency of surgeries. However there are many unsolved problems that can be investigated in future research, of which a selection is listed below:

- Automated assessment of a surgeon's skill during surgery: It is an unresolved problem of how to measure surgical skills (e.g. motoric skills) in-vivo in surgery. There are a lot of signals that can be used to measure surgical tasks, however, how to analyse those signals to give an objective assessment about surgical skills is currently unresolved.

- Extending the results to other surgical specialties: To the best of the author's knowledge, automatic workflow segementation is mostly applied in laparoscopic cholecystectomy because of its standard workflow and high frequency of occurrence. We recommend exploring the possibilities for workflow segmentation for other specialties of surgery.

- Continuous online updating of predicted parameters: This thesis shows how prior knowledge can be used to predict parameters. For future work, we propose to continuously update these estimates during surgery to get an increasingly better estimate as the procedure nears completion.

- Workflow logging for novel surgical techniques: It is hard to predict the future of surgical procedures. After more than 20 years since its introduction, laparoscopic surgery is steadily becoming the conventional technique. Today, new surgical treatment methods are introduced such as monoport, NOTES (natural port) and robotic surgery. Robotic surgery is currently in its infancy and faces many obstacles, yet could produce huge amounts of data about kinematics of surgical movements. We propose to explore the low-level data produced from (robotic) surgery to improve the accuracy of workflow logs for surgical procedures.

Appendix A

Supplementary material for Chapter 2

In this appendix we show how to evaluate the performance of classifiers.

Classification performance criterion

Assume a classifier is constructed (i.e. trained) using known instances of patients belonging to one of two classes. Although the classifier can show very good performance for all patients in the training dataset, this is no guarantee that the classifier will perform well for new datapionts (See Figure 2.2 for an illustration of this). Thus the performance of the classifiers needs to be evaluated with new datapoints.

Different criteria can be used to evaluate the classification performance. Common criteria to evaluate classification performance are the classification error and the Area Under the Curve (AUC) of an (Receiver Operating Curve) ROC curve. The classification error is often used for classifier evaluation through the straightforward counting and calculation of the percentage of misclassified records in a test set. For our two class classification problem, assume that a classifier f is evaluated on a test set $\chi = \{(x_i, y_i), i = 1, ..., N\}$, with $x_i \in \mathbb{R}^p$ as the p-dimensional feature vectors and $y \in \{\omega^+, \omega^-\}$ as class labels indicating whether surgeries are easy or complex. The classification error is estimated by: $\epsilon = \frac{1}{N}\sum_{i=1}^{N} I(f(x) \neq y_i)$, where $I(.)$ is the indicator function that outputs 1 when the statement is true and 0 otherwise [Tax08]. One disadvantage to this measure is that it is sensitive to class priors [Tax08].

The AUC is a natural criterion for measuring the classification performance of a classifier. In basic terms, it calculates the probability that a randomly selected positive (datapoint from the first class) is ranked before a randomly selected negative (datapoint from the second class) [Lok01]. It is a widely used measure of ranking performance. It can be calculated by $E = 1 - AUC =$

$1 - \frac{1}{N^+N^-} \sum_{i=1}^{N+} \sum_{j=1}^{N-} I(f(x_i) > f(x_j))$, where N^+ and N^- refer to the number of objects from the positive and negative classes, respectively. For our dataset, where $N^+ = 257$ and $N^- = 80$, the AUC error remains a relative measure independent of those priors. Furthermore, the AUC tends to generate a more stable estimate of performance than does the classification error [Tax08].

Classifier learning curve

The performance of classification algorithms often increases with the number of observations used to train the algorithm. Also if we have an infinitely large set of training examples the most complex classifier will generally have a better performance when compared to a simple classifier. For small sample sizes, however, simple classifiers are preferred since complex classifiers may overfit the pattern as described in Section 2.3.2. A learning curve is a graphical representation of this trade-off and is used to fine-tune the classifier complexity and appropriate training size. The learning curve plots the criteria (e.g. ϵ or AUC see SectionA) of the classifier for increasing training set sizes. The learning curve indicates per classifier what is a sufficiently large training set size and shows for each classifier the potential for performance improvement through the availability of more data. Figure A.1 shows the '*learning curve*' of two different classifiers, $Classifier1$ shows a flat learning curve after n patients. At n patients the classifier achieved its maximum performance and more training data would most likely not significantly improve the performance of the classifier. In contrast, the performance of $Classifier2$ is still improving which means it will most likely benefit further from more training data.

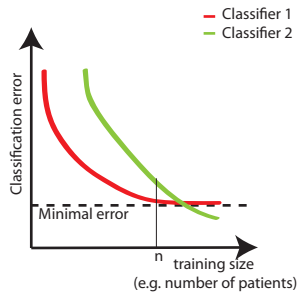


Figure A.1: *Classifier learning curve: $Classifier1$ shows a flat curve after n training samples, $Classifier2$ could use more training data to minimize its error.*

Feature learning curves

In many classification problems features are reduced to a smaller set to allow stable and better results. Feature selection is used as the first step for classific-

ation efficiency reasons to remove redundant features. Feature learning curves visualize the relationship between the training size, classification error and the number of features used (i.e dimensionality of the feature space) as illustrated in A.2. They are used to select the smallest number of features required, without negatively impacting the classification error. For high dimensional spaces we need more training samples to train a classifier and a more complex classifier to fit the pattern. For a finite training size, however, there is a risk that those complex classifiers overfit as described in Section 2.3.2, reducing the classification performance.

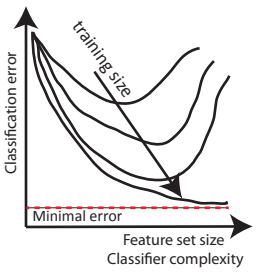


Figure A.2: *Feature learning curve: Given the number of training samples n, a minimum error is obtained for a specific number of features*

Feature ranking

For clinical studies, it is common to measure the significance and rank the individual features. Feature selection allows the selection of a subset of features proven to be significant in predicting a specific outcome. Individual features, when used alone, may not be relevant in predicting specific outcome, while in combination with other feature(s) they make a strong predictor. Feature ranking is useful for determining the (clinically) relevant subset of features amongst all collected features.

By randomly selecting a given percentage of datapoints from the training set, the features can be ranked using a feature selection mechanism. By repeating this experiment many times, insight can be obtained in the significance and the stability of a specific subset of features. Note that even the same features could be ranked in different orders, depending upon the samples selected from the training set. Figure A.3 shows an example of a ranking plot of different features from 350 runs. For each run, let's say 50% of the samples were randomly selected from the dataset and ranked by using a feature selection algorithm [Tax08]. The general trend is for each rank level to be dominated by a small number of features. Exception to this trend appeared in the higher levels (8-10), where the features become apparently less instable and hence less relevant for the classification task.

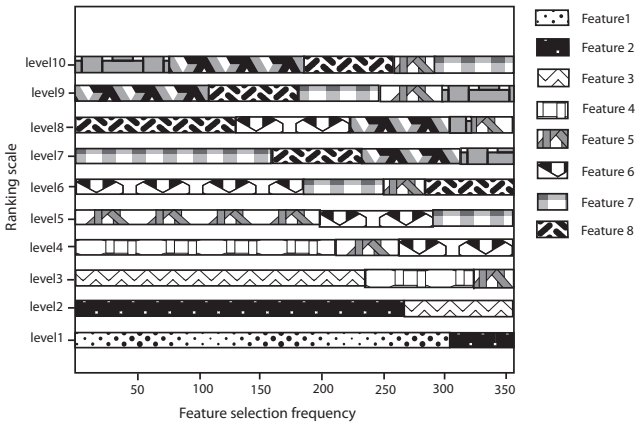


Figure A.3: *Feature ranking using a specific feature selection mechanism*

Appendix B

Supplementary material for Chapter 7

In this appendix describe the pre-operative features used for predicting the post anaesthesia care unit - length of stay (PACU LOS)

B.1 Pre-operative features

- Gender: Binary feature referring to patient Gender
- Age: Ordinal feature referring to patient Age
- BMI: Ordinal feature referring to Body Mass Index
- Current Smoker: Binary feature referring to Current Smoker
- HTN: Binary feature referring to Hypertension Diagnosis
- Diabetes: Binary feature referring to Diabetes diagnosis
- Cardiac Failure: Binary feature referring to Cardiac Failure diagnosis
- COPD: Binary feature referring to COPD diagnosis
- Cocaine Use Current: Binary referring to Cocaine Use Current
- Cocaine Use Past: Binary feature referring to Cocaine Use Past
- Alcohol Use Current: Binary feature referring to Alcohol Use Current
- Alcohol Use Past: Binary feature referring to Alcohol Use Past
- Opiate Use Current: Binary feature referring Opiate Use Current

- Opiate Use Past: Binary feature referring Opiate Use Past
- Anaesthetists ID: Binary feature referring Anaesthetists ID
- Surgeon ID: Binary feature referring Surgeon ID
- Anes Type General: Binary feature referring Primary Anesthetic - General Anaesthesia
- Anes Type MAC: Binary feature referring Primary Anesthetic - Monitored Anaesthesia Care
- Anes Type Regional: Binary feature referring to Primary Anesthetic - Regional Anaesthesia
- Anes Attending: Binary feature referring to Anaesthesia Attending Name
- ASA: Discreet feature referring to ASA physical status

B.2 Intra-operative features

- Anes: Resident Binary feature referring to Anaesthesia Resident Name
- CRNA: Binary feature referring to Certified Registered Nurse Anaesthetist Name
- SRNA: Binary feature referring to Student Registered Nurse Anaesthetist Name
- Handoff: Binary feature referring to transition in care by AnesAtteding, AnesResident, CRNA, SRNA
- Isoflurane: Used Binary feature referring to If Isoflurane is used before start of surgery
- Desflurane: Used Binary feature referring to If Desflurane is used before start of surgery
- Sevoflurane: Used Binary feature referring to If Sevoflurane is used before start of surgery
- Midazolam: Binary feature referring to If Midazolam is used before start of surgery
- Fentanyl: Binary feature referring to If Fentanyl is used before start of surgery
- Morphine: Binary feature referring to If Morphine is used before start of surgery

- Hydromorphone: Binary feature referring to If Hydromorphone is used before start of surgery
- Meperidine: Binary feature referring to If Meperidine is used before start of surgery
- Droperidol: Binary feature referring to If Droperidol is used before start of surgery
- Haloperidol: Binary feature referring to If Haloperidol is used before start of surgery
- Ondansetron: Binary feature referring to If Ondansetron is used before start of surgery
- Dexamethasone: Binary feature referring to If Dexamethasone is before start of surgery
- Scopalamine: Binary feature referring to If Scopalamine is used before start of surgery
- Neostigmine: Binary feature referring to If Neostigmine is used before start of surgery
- Physostigmine: Binary feature referring to If Physostigmine is used before start of surgery
- Glycopyrrollate: Binary feature referring to If Glycopyrrollate is used before start of surgery
- Phenergan: Binary feature referring to If Phenergan is used before start of surgery
- Ephedrine: Binary feature referring to If Ephedrine is used before start of surgery
- Phenylephrine: Binary feature referring to If Phenylephrine is used before start of surgery
- Dopamine: Binary feature referring to If Dopamine is used before start of surgery
- Epinephrine: Binary feature referring to If Epinephrine is used before start of surgery
- Dobutamine: Binary feature referring to If Dobutamine is used before start of surgery
- Vasopressin: Binary feature referring to If Vasopressin is used before start of surgery

- Ephedrine-AS: Binary feature referring to If Ephedrine is used before start of surgery
- Epinephrine-AS: Binary feature referring to If Epinephrine is used after start of surgery
- Phenylephrin-AS: Binary feature referring to If Phenylephrin is used after start of surgery
- Dopamine-AS: Binary feature referring to: If Dopamine is used after start of surgery
- Dobutamine-AS: Binary feature referring to: If Dobutamine is used after start of surgery
- Vasopressin-AS: Binary feature referring to Vasopressin used after start of surgery
- PRBC: Binary feature referring to packed red blood cells
- AlineUsedInCase: Binary feature referring to placing intra-arterial catheter
- CentralLineUsedInCase: Binary feature referring to Placing an Central venous line

B.3 Post-operative features

- Operative time: Ordinal feature referring to Perioperative time
- Anaesthesia time: Ordinal feature referring to Anaesthesia time
- Surgical time: Ordinal feature referring to Intraoperative time
- PONV: Binary feature referring to Postoperative Nausea and Vomiting

Bibliography

[Aal02] W. van der Aalst and K. van Hee. *Workflow Management: Models, Methods, and Systems*. MIT Press, 2002.

[Aal03] W. van der Aalst, B. van Dongen, J. Herbst, L. Maruster, G. Schimm, and A. Weijters. Workflow mining: A survey of issues and approaches. *Data & Knowledge Engineering*, vol. 47(2):pp. 237–267, 2003.

[Ada96] D. H. Adams and B. B. Abernathy. Laser ureterolithotripsy for cystine calculi. *AORN*, vol. 64(6):pp. 924–930, 1996.

[Ahm06] S. Ahmadi, T. Sielhorst, R. Stauder, M. Horn, H. Feussner, and N. Navab. Recovery of surgical workflow without explicit models. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*, pp. 420–428, 2006.

[Alb02] R. Albin. Sham surgery controls: intracerebral grafting of fetal tissue for parkinsons disease and proposed criteria for use of sham surgery controls. *Journal of medical ethics*, vol. 28(5):p. 322, 2002.

[Bea09] R. Beasley and R. Howe. Increasing accuracy in image-guided robotic surgery through tip tracking and model-based flexion correction. *Robotics, IEEE Transactions on*, vol. 25(2):pp. 292–302, 2009.

[Bis94] C. Bishop. Novelty detection and neural network validation. vol. 141(4):pp. 217–222, 1994.

[Bis95] C. Bishop. Neural networks for pattern recognition, 1995.

[Bla96] N. Black. Why we need observational studies to evaluate the effectiveness of health care. *Bmj*, vol. 312(7040):p. 1215, 1996.

[Blu08] T. Blum, N. Padoy, H. Feußner, and N. Navab. Workflow mining for visualization and analysis of surgeries. *International journal of computer assisted radiology and surgery*, vol. 3(5):pp. 379–386, 2008.

[Bou09] L. Bouarfa, P. Jonker, and J. Dankelman. Surgical context discovery by monitoring low-level activities in the OR. 2009.

[Bou10] L. Bouarfa, P. Jonker, and J. Dankelman. Discovery of high-level tasks in the operating room. *Journal of Biomedical Informatics*, 2010.

[Bou11a] L. Bouarfa, O. Akman, A. Schneider, P. P. Jonker, and J. Dankelman. In-vivo real-time tracking of surgical instruments in endoscopic video. *Minimally Invasive Therapy & Allied Technologies*, vol. 0(0):pp. 1–6, 2011.

[Bou11b] L. Bouarfa, P. Jonker, and J. Dankelman. Discovery of high-level tasks in the operating room. *Journal of Biomedical Informatics*, vol. 44(3):pp. 455 – 462, 2011. ¡ce:title¿Biomedical Complexity and Error¡/ce:title¿.

[Bou11c] L. Bouarfa, A. Schneider, H. Feussner, N. Navab, H. U. Lemke, P. P. Jonker, and J. Dankelman. Prediction of intraoperative complexity from preoperative patient data for laparoscopic cholecystectomy. *Artificial Intelligence in Medicine*, vol. 52(3):pp. 169 – 176, 2011.

[Bou12a] D. Bouarfa and J. Dankelman. Measuring consensus and detecting outliers from surgical process logs. *Submitted article*, 2012.

[Bou12b] L. Bouarfa, D. Tax, and J. Dankelman. Pattern recognition: a new perspective for evidence based surgery. *Submitted article*, 2012.

[Bou12c] L. Bouarfa, D. Tax, E. J.M., B. Rothman, and J. Dankelman. Length of stay in the post anaesthesia care unit - can it be estimated? *Submitted article*, 2012.

[Bro08] I. Brown, W. Jellish, B. Kleinman, E. Fluder, K. Sawicki, J. Katsaros, and R. Rahman. Use of postanesthesia discharge criteria to reduce discharge delays for inpatients in the postanesthesia care unit. *Journal of clinical anesthesia*, vol. 20(3):pp. 175–179, 2008.

[Car10a] B. Cardoen, E. Demeulemeester, and J. Beliën. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, vol. 201(3):pp. 921–932, 2010.

[Car10b] B. Cardoen, E. Demeulemeester, and J. Belin. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, vol. 201(3):pp. 921 – 932, 2010.

[Chu99] F. Chung and G. Mezei. Factors contributing to a prolonged stay after ambulatory surgery. *Anesthesia & Analgesia*, vol. 89(6):p. 1352, 1999.

[Coh99] M. Cohen, L. O'Brien-Pallas, C. Copplestone, R. Wall, J. Porter, and K. Rose. Nursing workload associated with adverse events in the postanesthesia care unit. *Anesthesiology*, vol. 91(6):p. 1882, 1999.

[Com02] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24(5):pp. 603–619, May 2002.

[Cox93] M. Cox, T. Wilson, A. Luck, P. Jeans, R. Padbury, and J. Toouli. Laparoscopic cholecystectomy for acute inflammation of the gallbladder. *Annals of surgery*, vol. 218(5):pp. 630–634, 1993.

[Dar07] P. Darcy, B. Stantic, R. Derakhshan, and D. Parsons. Correcting Stored RFID Data with Non-Monotonic Reasoning. *international journal of principles and applications of information science and technology*, vol. 1(1), December 2007.

[Dav05] D. Davenport, W. Henderson, S. Khuri, and R. Mentzer Jr. Preoperative risk factors and surgical complexity are more predictive of costs than postoperative complications: a case study using the National Surgical Quality Improvement Program (NSQIP) database. *Annals of surgery*, vol. 242(4):pp. 263–471, 2005.

[DB07] M. De Bruijne, M. Zegers, and C. Hoonhout, L.H.F.and Wagner. Onbedoelde schade in nederlandse ziekenhuizen: Dossieronderzoek van ziekenhuizenopnames in 2004. 2007.

[Dek01] W. Dekkers and G. Boer. Sham neurosurgery in patients with parkinson's disease: is it morally acceptable? *Journal of medical ethics*, vol. 27(3):pp. 151–156, 2001.

[Dex95a] F. Dexter and J. Tinker. Analysis of strategies to decrease postanesthesia care unit costs. *Anesthesiology*, vol. 82(1):p. 94, 1995.

[Dex95b] F. Dexter and J. Tinker. Analysis of strategies to decrease postanesthesia care unit costs. *Anesthesiology*, vol. 82(1):p. 94, 1995.

[Dex99] F. Dexter, A. Macario, P. Manberg, and D. Lubarsky. Computer simulation to determine how rapid anesthetic recovery protocols to decrease the time for emergence or increase the phase i postanesthesia care unit bypass rate affect staffing of an ambulatory surgery center. *Anesthesia & Analgesia*, vol. 88(5):p. 1053, 1999.

[Dex05] F. Dexter, R. Epstein, E. Marcon, and R. de Matta. Strategies to reduce delays in admission into a postanesthesia care unit from operating rooms. *Journal of PeriAnesthesia Nursing*, vol. 20(2):pp. 92–102, 2005.

[Dex06] F. Dexter, M. Davis, C. E. Halbeis, R. Marjamaa, J. Marty, C. McIntosh, Y. Nakata, K. N. Thenuwara, T. Sawa, and M. Vigoda. Mean operating room times differ by 50% among hospitals in different countries for laparoscopic cholecystectomy and lung lobectomy. *Journal of Anesthesia*, vol. 20:pp. 319–322, 2006.

[Dim04] J. B. Dimick, S. L. Chen, P. A. Taheri, W. G. Henderson, S. F. Khuri, D. A. Campbell, and Jr. Hospital costs associated with surgical complications: A report from the private-sector national surgical quality improvement program. *Journal of the American College of Surgeons*, vol. 199(4):pp. 531 – 537, 2004.

[Din00] H.-P. Dinkel, S. Kraus, J. Heimbucher, R. Moll, J. Knupffer, H.-J. Gassel, S. M. Freys, K.-H. Fuchs, and G. Schindler. Sonography for selecting candidates for laparoscopic cholecystectomy: A prospective study. *American Journal of Roentgenology*, vol. 174(5):pp. 1433–1439, 2000.

[Din04] D. Dindo, N. Demartines, and P. Clavien. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Annals of surgery*, vol. 240(2):p. 205, 2004.

[Doe09] R. Does, T. Vermaat, J. Verver, S. Bisgaard, and J. Van Den Heuvel. Reducing start time delays in operating rooms. *Journal of Quality Technology*, vol. 41(1):pp. 95–109, 2009.

[Dui07a] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. De Ridder, D. Tax, and S. Verzakov. A matlab toolbox for pattern recognition. Delft University of Technology, The Netherlands, pp. 1–61. Prtools 4 edn., 2007. Toolbox manual available at:`http://www.prtools.org/files/PRTools4.1.pdf`.

[Dui07b] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. De Ridder, D. Tax, and S. Verzakov. A matlab toolbox for pattern recognition, 2007. Toolbox manual available at:`http://www.prtools.org/files/PRTools4.1.pdf`.

[Eck10] S. Eckermann, J. Karnon, and A. Willan. The value of value of information: best informing research design and prioritization using current methods. *PharmacoEconomics*, vol. 28(9):pp. 699–709, 2010.

[Eng05] D. Engels. On Ghost Reads in RFID Systems, 2005.

[Far10] F. Farrokhyar, P. Karanicolas, A. Thoma, M. Simunovic, M. Bhandari, P. Devereaux, M. Anvari, A. Adili, and G. Guyatt. Randomized controlled trials of surgical interventions. *Annals of surgery*, vol. 251(3):p. 409, 2010.

[Fre87] B. Freedman. Equipoise and the ethics of clinical research. *New England Journal of Medicine*, vol. 317(3):pp. 141–145, 1987.

[Fuk90] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Pr, 1990.

[Gab09] R. Gabriel, S. Kumar, and A. Shrestha. Evaluation of predictive factors for conversion of laparoscopic cholecystectomy. *Kathmandu University Medical Journal*, vol. 7(1):pp. 26–30, 2009.

[Geo08] C. Georgiades, T. Mavromatis, G. Kourlaba, S. Kapiris, E. Bairamides, A. Spyrou, C. Kokkinos, C. Spyratou, M. Ieronymou, and G. Diamantopoulos. Is inflammation a significant predictor of bile duct injury during laparoscopic cholecystectomy? *Surgical Endoscopy*, vol. 22:pp. 1959–1964, 2008.

[GHO08] Z. GHOGAWALA, F. BARKER, and B. CARTER. Clinical equipoise and the surgical randomized controlled trial. *Neurosurgery*, vol. 62(6):p. N9, 2008.

[GM10] S. Gonzlez-Moreno. Standardisation and outcomes audit: a step forward in surgical oncology. *Clinical and Translational Oncology*, vol. 12:pp. 389–390, 2010. 10.1007/s12094-010-0523-7.

[Gon07] H. Gonzalez, J. Han, and X. Shen. Cost-conscious cleaning of massive RFID data sets. In *IEEE 23rd International Conference on Data Engineering, 2007. ICDE 2007*, pp. 1268–1272. 2007.

[Gri08] F. Griffin and D. Classen. Detection of adverse events in surgical patients using the trigger tool approach. *Quality and Safety in Health Care*, vol. 17(4):p. 253, 2008.

[Gro07] J. Groopman and M. Prichard. *How doctors think*. Springer, 2007.

[H.00] G. H. and Eltabbakh. Effect of surgeon's experience on the surgical outcome of laparoscopic surgery for women with endometrial cancer. *Gynecologic Oncology*, vol. 78(1):pp. 58 – 61, 2000.

[Hal05] S. Hall, C. DArcy, J. Holman, J. Finn, and J. Semmens. Improving the evidence base for promoting quality and equity of surgical care using population-based linkage of administrative health records. *International Journal for Quality in Health Care*, vol. 17(5):p. 415, 2005.

[Hei04] F. van der Heijden, R. Duin, D. De Ridder, and D. Tax. *Classification, parameter estimation and state estimation*. Wiley Online Library, 2004.

[Hod04a] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, vol. 22(2):pp. 85–126, 2004.

[Hod04b] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, vol. 22(2):pp. 85–126, 2004.

[Hoo09] L. Hoonhout, M. de Bruijne, C. Wagner, M. Zegers, R. Waaijman, P. Spreeuwenberg, H. Asscheman, G. Van Der Wal, and M. van Tulder. Direct medical costs of adverse events in dutch hospitals. *BMC health services research*, vol. 9(1):p. 27, 2009.

[Hoz05] W. Hozack. Optimizing clinical performance–the rothman institute joint replacement service at thomas jefferson university hospital. *Health Policy Newsletter*, vol. 13(2):p. 3, 2005.

[Hu08] D. Hu, S. Pan, V. Zheng, N. Liu, and Q. Yang. Real world activity recognition with multiple goals. In *Proceedings of the 10th international conference on Ubiquitous computing*, pp. 30–39. ACM New York, NY, USA, 2008.

[Jam07] A. James, D. Vieira, B. Lo, A. Darzi, and G. Yang. Eye-gaze driven surgical workflow segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007*, pp. 110–117, 2007.

[JCB10] R. Jagadeesh Chandra Bose and W. van der Aalst. Trace alignment in process mining: opportunities for process diagnostics. *Business Process Management*, pp. 227–242, 2010.

[Jea10] A. Jeang and A. Chiang. Economic and quality scheduling for effective utilization of operating rooms. *Journal of Medical Systems*, pp. 1–18, 2010.

[Jen01] F. Jensen. *Bayesian Networks and Decision Graphs*. Springer, 2001.

[Jen10] E. Jenkins, V. Yom, L. Melman, R. Pierce, R. Schuessler, M. Frisella, J. Christopher Eagon, L. Michael Brunt, and B. Matthews. Clinical predictors of operative complexity in laparoscopic ventral hernia repair: a prospective study. *Surgical Endoscopy*, vol. 24:pp. 1872–1877, 2010.

[Joh94] A. Johnson. Surgery as a placebo. *The Lancet*, vol. 344(4):pp. 1140–1142, 1994.

[Joh05] T. Johnston and A. Schembri. The use of ELAN annotation software in the Auslan Archive/Corpus Project. In *Presentation at the Ethnographic Eresearch Annotation Conference, University of Melbourne*. 2005.

[Kno08] J. Knottnerus and F. Buntinx. *The evidence base of clinical diagnosis*. Wiley Online Library, 2008.

[Kra05] J. Kral, J. Dixon, F. Horber, S. Rssner, S. Stiles, J. Torgerson, and H. Sugerman. Flaws in methods of evidence-based medicine may adversely affect public health directives. *Surgery*, vol. 137(3):pp. 279 – 284, 2005.

[Lee10] M. C. Lee, L. Boroczky, K. Sungur-Stasik, A. D. Cann, A. C. Borczuk, S. M. Kawut, and C. A. Powell. Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction. *Artificial Intelligence in Medicine*, vol. 50(1):pp. 43 – 53, 2010. Knowledge Discovery and Computer-Based Decision Support in Biomedicine.

[Lew06] M. C. Lewis, R. I. Gerenstein, and G. Chidiac. Onset time for sevoflurane/nitrous oxide induction in adults is prolonged with increasing age. *Anesthesia & Analgesia*, vol. 102(6):pp. 1699–1702, 2006.

[Lin09] R.-H. Lin. An intelligent model for liver disease diagnosis. *Artificial Intelligence in Medicine*, vol. 47(1):pp. 53 – 62, 2009.

[Lo03] B. Lo, A. Darzi, and G. Yang. Episode classification for the analysis of tissue/instrument interaction with multiple visual cues. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003*, pp. 230–237, 2003.

[Lo07] B. Lo. *Activity Profiling for Minimally Invasive Surgery*. Ph.D. thesis, Imperial College London, 2007.

[Lok01] Y. Loke and S. Derry. Reporting of adverse drug reactions in randomised controlled trials–a systematic survey. *BMC Clinical Pharmacology*, vol. 1(1):p. 3, 2001.

[Mac95] A. Macario, T. Vitez, B. Dunn, and T. McDonald. Where are the costs in perioperative care?: Analysis of hospital costs and charges for inpatient surgical care. *Anesthesiology*, vol. 83(6):p. 1138, 1995.

[Mac99] R. Macklin. The ethical problems with sham surgery in clinical research. *New England Journal of Medicine*, vol. 341(13):pp. 992–996, 1999.

[Mai98] J. Maintz and M. A. Viergever. A survey of medical image registration. *Medical Image Analysis*, vol. 2(1):pp. 1 – 36, 1998.

[Mar03] E. Marcon, S. Kharraja, N. Smolski, B. Luquet, and J. Viale. Deterf the number of beds in the postanesthesia care unit: a computer simulation flow approach. *Anesthesia & Analgesia*, vol. 96(5):p. 1415, 2003.

[Mar06] E. Marcon and F. Dexter. Impact of surgical sequencing on post anesthesia care unit staffing. *Health care management science*, vol. 9(1):pp. 87–98, 2006.

[McK95] J. McKernan and J. Champion. Access techniques: Veress needle–initial blind trocar insertion versus open laparoscopy with the Hasson trocar. *Endoscopic surgery and allied technologies*, vol. 3(1):pp. 35–38, 1995.

[McL99] R. S. McLeod. Issues in surgical randomized controlled trials. *World Journal of Surgery*, vol. 23:pp. 1210–1214, 1999. 10.1007/s002689900649.

[Meg06a] G. Megali, S. Sinigaglia, O. Tonet, F. Cavallo, and P. Dario. Understanding expertise in surgical gesture by means of Hidden Markov Models. In *Biomedical Robotics and Biomechatronics, 2006. BioRob 2006. The First IEEE/RAS-EMBS International Conference on*, pp. 625–630. IEEE, 2006.

[Meg06b] G. Megali, S. Sinigaglia, O. Tonet, and P. Dario. Modelling and evaluation of surgical performance using hidden Markov models. *Biomedical Engineering, IEEE Transactions on*, vol. 53(10):pp. 1911–1919, 2006.

[Mil03] F. Miller. Sham surgery: An ethical analysis. *American Journal of Bioethics*, vol. 3(4):pp. 41–48, 2003.

[Miy09] F. Miyawaki, T. Tsunoi, H. Namiki, T. Yaginuma, K. Yoshimitsu, D. Hashimoto, and Y. Fukui. Development of automatic acquisition system of surgical-instrument informantion in endoscopic and laparoscopic surgey. In *Industrial Electronics and Applications, 2009. ICIEA 2009. 4th IEEE Conference on*, pp. 3058 –3063. May 2009.

[Mur04] G. Murphy, J. Szokol, M. Franklin, J. Marymont, M. Avram, and J. Vender. Postanesthesia care unit recovery times and neuromuscular blocking drugs: A prospective study of orthopedic surgical patients randomized to receive pancuronium or rocuronium. *Anesthesia & Analgesia*, vol. 98(1):p. 193, 2004.

[Neu06a] T. Neumuth, N. Durstewitz, M. Fischer, G. Strauß, A. Dietz, J. Meixensberger, P. Jannin, K. Cleary, H. Lemke, and O. Burgert. Structured recording of intraoperative surgical workflows. In *Proceedings of SPIE*, vol. 6145, p. 61450A. 2006.

[Neu06b] T. Neumuth, G. Strauß, J. Meixensberger, H. Lemke, and O. Burgert. Acquisition of process descriptions from surgical interventions. In *Database and expert systems applications*, pp. 602–611. Springer, 2006.

[Neu08] T. Neumuth, S. Mansmann, M. Scholl, and O. Burgert. Data warehousing technology for surgical workflow analysis. In *Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on*, pp. 230–235. IEEE, 2008.

[Noc10] P. Nocini, G. Verlato, A. Frustaci, A. de Gemmis, G. Rigoni, and D. De Santis. Evidence-based dentistry in oral surgery: Could we do better? *The open dentistry journal*, vol. 4:p. 77, 2010.

[Nor03] J. A. Norton. *Essential practice of surgery: basic science and clinical evidence*. Springer Verlag, 2003.

[ope] http://opencv.willowgarage.com/.

[Ost10] D. J. Ostlie and S. D. S. Peter. The current state of evidence-based pediatric surgery. *Journal of Pediatric Surgery*, vol. 45(10):pp. 1940 – 1946, 2010.

[Pad07] N. Padoy, T. Blum, I. Essa, H. Feussner, M. Berger, and N. Navab. A boosted segmentation method for surgical workflow analysis. In *Proceedings of the 10th international conference on Medical image computing and computer-assisted intervention-Volume Part I*, pp. 102–109. Springer-Verlag, 2007.

[Pad08] N. Padoy, T. Blum, H. Feußner, M. Berger, and N. Navab. On-line recognition of surgical activity for monitoring in the operating room. In *Proceedings of the 20th Conference on Innovative Applications of Artificial Intelligence (IAAI-08)*. 2008.

[Pad10] N. Padoy, T. Blum, S. Ahmadi, H. Feussner, M. Berger, and N. Navab. Statistical modeling and recognition of surgical workflow. *Medical Image Analysis*, 2010.

[Pan07] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human computing and machine understanding of human behavior: A survey. *Lecture Notes in Computer Science*, vol. 4451:p. 47, 2007.

[Pat99] V. Patel, J. Arocha, and D. Kaufman. Expertise and tacit knowledge in medicine. *Tacit knowledge in professional practice: Researcher and practitioner perspectives*, pp. 75–99, 1999.

[Per01] R. Perugini and M. Callery. Complications of laparoscopic surgery. *Surgical Treatment: Evidence-Based and Problem-Oriented*, 2001.

[Per08] J. Perlmutter. Understanding clinical trial design: A tutorial for research advocates. *Research Advocacy Network's Advocate Institute*, 2008.

[Pha07] V. Pham, Q. Qiu, A. Wai, and J. Biswas. Application of ultrasonic sensors in a smart environment. *Pervasive and Mobile Computing*, vol. 3(2):pp. 180–207, 2007.

[Rab89] L. Rabiner et al. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, vol. 77(2):pp. 257–286, 1989.

[Ree07] B. Rees. Learning surgery: The surgery clerkship manual. *Annals of The Royal College of Surgeons of England*, vol. 89(3):p. 334, 2007.

[Rev90] J. Reves and L. Smith. From monitoring to predicting outcome. *Annals of surgery*, vol. 212(5):p. 559, 1990.

[Riv08] N. Rivera, R. Mountain, L. Assumpcao, A. Williams, A. Cooper, D. Lewis, R. Benson, J. Miragliotta, M. Marohn, and R. Taylor. ASSIST-Automated System for Surgical Instrument and Sponge Tracking. In *RFID, 2008 IEEE International Conference on*, pp. 297–302. IEEE, 2008.

[Sam06] K. Samad, M. Khan, F. Hameedullah, M. Hamid, and F. Khan. Unplanned prolonged postanaesthesia care unit length of stay and factors affecting it. *JPMA*, vol. 56(108), 2006.

[Sch09] T. Schoenmeyr, P. Dunn, D. Gamarnik, R. Levi, D. Berger, B. Daily, W. Levine, and W. Sandberg. A model for understanding the impacts of demand and capacity on waiting time to enter a congested recovery room. *Anesthesiology*, vol. 110(6):p. 1293, 2009.

[Sho06] E. Shortliffe, J. Cimino, and I. NetLibrary. *Biomedical informatics: computer applications in health care and biomedicine*. Springer, 2006.

[Sli95] K. Slim, D. Pezet, J. Stencl, C. Lechner, S. Roux, P. Lointier, and J. Chipponi. Laparoscopic cholecystectomy: an original three-trocar technique. *World journal of surgery*, vol. 19(3):pp. 394–397, 1995.

[Sod10] M. Sodergren, F. Orihuela-Espina, J. Clark, J. Teare, G. Yang, and A. Darzi. Evaluation of Orientation Strategies in Laparoscopic Cholecystectomy. *Annals of surgery*, vol. 252(6):pp. 1027–1036, 2010.

[Sol95] M. Solomon and R. McLeod. Should we be performing more randomized controlled trials evaluating surgical operations?*. *Surgery*, vol. 118(3):pp. 459–467, 1995.

[Sta10] C. Staub, A. Knoll, T. Osa, and R. Bauernschmitt. Autonomous high precision positioning of surgical instruments in robot-assisted minimally invasive surgery under visual guidance. In *2010 Sixth International Conference on Autonomic and Autonomous Systems*, pp. 64–69. IEEE, 2010.

[Sto06] J. Stoll, P. Novotny, R. Howe, and P. Dupont. Real-time 3d ultrasound-based servoing of a surgical instrument. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pp. 613–618. IEEE, 2006.

[Tal11] N. Taleb. *The black swan: The impact of the highly improbable*. Random House Inc, 2011.

[Tap04] E. Tapia, S. Intille, and K. Larson. Activity recognition in the home using simple and ubiquitous sensors. *Lecture Notes in Computer Science*, pp. 158–175, 2004.

[Tax98] D. Tax and R. Duin. Outlier detection using classifier instability. *Advances in Pattern Recognition*, pp. 593–601, 1998.

[Tax08] D. Tax and R. Duin. Learning Curves for the Analysis of Multiple Instance Classifiers, 2008.

[Tes99] M. Tessler, L. Mitmaker, R. Wahba, and C. Covert. Patient flow in the post anesthesia care unit: an observational study. *Canadian Journal of Anesthesia / Journal canadien d'anesthsie*, vol. 46:pp. 348–351, 1999. 10.1007/BF03013226.

[The09] S. Theodoridis, K. Koutroumbas, A. Pikrakis, and D. Cavouras. *Introduction to pattern recognition: a matlab approach*. Academic Pr, 2009.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, pp. 267–288, 1996.

[Twe08] R. Twersky and B. Philip. *Handbook of ambulatory anesthesia*. Springer Verlag, 2008.

[UK12a] UK. postoperative complications after surgery, 2012.

[UK12b] UK. Rates of survival after heart surgery in the uk, 2012.

[Vak09] M. Vaknkipuram, K. Kahol, G. Islam, T. Cohen, and V. Patel. Visualization and analysis of medical errors in immersive virtual environments. *proceedings from the 17 th Annual Medicine Meets Virtual Reality Conference, MMVR2009,*, jan 19th-22th 2009.

[Wad98] J. Waddle, A. Evers, and J. Piccirillo. Postanesthesia care unit length of stay: quantifying and assessing dependent factors. *Anesthesia & Analgesia*, vol. 87(3):p. 628, 1998.

[Wal11] J. Walker. Obstetric statistics. *Dewhurst's Textbook of Obstetrics & Gynaecology*, pp. 394–408, 2011.

[Wan09] S. Wang, X. Yin, B. Ge, Y. Gao, H. Xie, and L. Han. Machine Vision for Automated Inspection of Surgical Instruments. In *Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on*, pp. 1–4. IEEE, 2009.

[Wel95] G. Welch and G. Bishop. An introduction to the kalman filter, 1995.

[Wol07] B. Wolff, J. Fleshman, and D. Beck. *The ASCRS textbook of colon and rectal surgery*. Springer Verlag, 2007.

[Zeg07] M. Zegers, M. de Bruijne, C. Wagner, P. Groenewegen, R. Waaijman, and G. Van Der Wal. Design of a retrospective patient record study on the occurrence of adverse events among patients in dutch hospitals. *BMC Health Services Research*, vol. 7(1):p. 27, 2007.

[Zen09] J. Zeng, J. Duan, and C. Wu. A new distance measure for hidden Markov models. *Expert Systems With Applications*, 2009.

[Zis96] A. Zisman, R. Gold-Deutch, E. Zisman, M. Negri, Z. Halpern, G. Lin, and A. Halevy. Is male gender a risk factor for conversion of laparoscopic into open cholecystectomy? *Surgical Endoscopy*, vol. 10:pp. 892–894, 1996.

Summary

Recognizing surgical patterns

In the Netherlands, each year over 1700 patients die from preventable surgical errors. Numerous initiatives to improve surgical practice have had some impact, but problems persist. Despite the introduction of checklists and protocols, patient safety in surgery remains a continuing challenge. This is complicated by some surgeons viewing their own work as an artistic manoeuver whose workflow cannot be captured. However, safeguarding patient safety is also a hospital's management responsibility and no longer only in the surgeon's hands.

In spite of the inherent variations, surgeries of the same kind produce similar data, and are usually performed in similar workflows. Surgery is characterized by a peri-operative pipeline of pre-, intra- and post-operative processes. To both reduce errors and improve efficiency, the workflow in the peri-operative pipeline should be designed and planned as effectively as possible in terms of flow of patients and allocation of scarce resources such as operating rooms, instruments and personnel. Currently, planning is done on a very basic level, without using real-world data to learn and improve efficiency. Fortunately, there is lot of available, but unexploited data about surgical interventions that can be used for this purpose.

The aim of this thesis is to use acquired and registered peri-operative data to support hospital management to improve safety and efficiency in surgery. The method of assessing safety and efficiency in surgery for individual patients needs to be tailored to each patient. As a result generalization of the results is difficult. We discuss how pattern recognition (PR) provides tools for the assessment of surgical outcome for individual patients. It also allows for handling of outliers and does not impose the same restrictions on data collection procedures as for randomized controlled trials. We show that PR is a pragmatic next step towards data intensive operating rooms with evidence based support for surgeries. Below the techniques as proposed in this thesis are briefly described.

To support pre-operative planning of surgeries, assessment of surgical complexity is needed beforehand in order to prepare and possibly avoid complications and

delays. This complexity assessment can also aid surgeons in decisions regarding how to proceed with the surgical procedure, for instance by taking extra precautions or making a referral to a more experienced surgeon when a complex surgery is predicted. We show how to use readily available patient data to predict surgical complexity. Classifiers are trained and evaluated using readily collected data from patients undergoing laparoscopic cholecystectomy (LAPCHOL). It is shown that complexity of LAPCHOL surgeries can be predicted in the pre-operative stage with an accuracy up to 83% using an LDC or SVM classifier. We also derived the set of features that are relevant for predicting complexity including inflammation, wall thickening, sex and BMI score.

To realize intra-operative safety and efficiency goals in surgery, hospitals are searching for autonomous systems for monitoring the surgical workflow in the operating room (OR). In this thesis we propose an autonomous registration technique for the OR. Registering the time of use of surgical instruments and the sequence in which they are used enables us to detect the surgical steps, including the duration of each step. By deploying this as a realtime system, dynamic support for the surgical team and dynamic planning of patients can be performed.

For monitoring the usage of surgical instruments, signals from sensors which can detect video, motion and RFID tags can be used. For the application in the OR, it is necessary that these sensors are designed to meet the requirements of the OR environment, specifically with respect to sterilization and non-intrusiveness. We propose a tracking system to detect and track instruments in endoscopic video using biocompatible and sterilization-proof colour markers. The system tracks single and multiple instruments in the video. The output of the tracking tool is a log file with an identifier of the instrument used and the duration of its use for each entry.

These instrument logs are then used for workflow mining and outlier detection in surgery. We derived a surgical consensus from multiple surgery logs using global multiple sequence alignment. We showed that the derived consensus conforms to the main steps of laparoscopic cholecystectomy as described in best practices. Using global pair-wise alignment, we showed that outliers from this consensus can be detected using the surgical log. These outliers are commonly simple variations in the execution of the surgical procedure, but can also represent serious complications or errors.

To improve post-operative efficiency, accurate predictions of patients' length of stay (LOS) in the postanesthesia care unit (PACU) may lead to cost savings and a number of other efficiency benefits. We propose to use available perioperative data to predict the PACU LOS, using the features case demographics, intra-operative parameters, medications, patient co-morbidities, and surgeon. A linear regression method was used along with ordinary least square regression and 'least absolute shrinkage and selection operator' (LASSO-) regression. A forward feature selection approach was then used to identify and rank factors that impact PACU LOS. We showed that PACU LOS can be predicted by perioperative factors with an improvement of 12-18 minutes compared to using the

mean baseline. If this prediction is updated with online information, mainly by monitoring post-operative oxygen saturation, future work could lead to real-time LOS algorithms based on peri-operative factors to predict, manage and possibly intercept anticipated, prolonged PACU LOS.

This thesis has proposed and demonstrated the application of pattern recognition tools to log, assess and predict surgical workflow parameters. Work in this thesis did not directly contribute to reduce errors and safety in the OR. However, the tools developed in the thesis can be used to support standardization of surgical workflow to both reduce errors and support surgical planning. Moreover, the proposed techniques for the operating room can be used in other medical domains such as the intensive care unit with only small contextual modifications.

Loubna Bouarfa

Samenvatting

Herkennen van patronen in chirurgie

In Nederland overlijden ieder jaar 1700 patiënten vanwege te vermijden fouten in de operatiekamer. Een veelvoud van initiatieven ter verbetering van de chirurgie hebben de situatie verbeterd, echter nog niet opgelost. Zelfs met de introductie van checklists en protocollen blijft het waarborgen van patiëntveiligheid voor, tijdens en na operaties een voortdurende strijd. Een complicerende factor is het beeld dat sommige chirurgen van het opereren hebben: meer een kunstvorm dan het volgen van een in een workflow vastgelegde processen en procedures. Echter, het waarborgen van patiëntveiligheid is nu ook een zaak van het bestuur van het ziekenhuis en niet alleen in de handen van de chirurg.

Ondanks de inherente verschillen in verschillende uitvoeringen van dezelfde chirurgische ingreep, leveren ze dezelfde soort gegevens op en worden meestal op een gelijksoortige manier uitgevoerd. Hiermee zijn ze met elkaar te vergelijken. Een chirurgische ingreep is op te delen in een opeenvolgende serie aan peri-operatieve processen in de pre-, intra- en post-operatieve fase. Om voor chirurgische ingrepen zowel fouten te vermijden als de efficiëntie te verhogen is het van belang deze serie van processen zo effectief mogelijk te ontwerpen en in te plannen zodat de doorstroming van patiënten en het toekenning van beperkt beschikbare middelen (o.a. instrumenten, locatie en mensen) wordt geoptimaliseerd. De huidige planningssystematiek in ziekenhuizen is meestal zeer basaal. Hier wordt er geen gebruik wordt gemaakt van praktijkgegevens om de systematiek bij te sturen en zo de efficiëntie vanuit de praktijk te verbeteren. Dit is een gemiste kans, aangezien de hiervoor benodigde gegevens vaak al wel worden verzameld, maar echter niet voor dit doel worden ingezet. Het doel van dit proefschrift is om bestaande, reeds verzamelde, gegevens rondom de operatie te gebruiken om het management van een ziekenhuis te ondersteunen in het verbeteren van de patiëntveiligheid en efficiëntie van chirurgische ingrepen.

De veiligheid en efficiëntie van een enkele chirurgische ingreep is zo sterk afhankelijk van de situatie van de patiënt dat het bepalen hiervan niet op een generieke

wijze kan plaatsvinden. Hierdoor is het zeer lastig conclusies te trekken die altijd van toepassing zijn en zijn er dus meestal zeer veel uitzonderingen op een enkele regel. Technieken van patroonherkenning (PR) bieden mogelijkheden om de uitkomst van een chirurgische ingreep op gebied van veiligheid en efficiëntie te beoordelen. PR biedt daarnaast een raamwerk voor het identificeren en analyseren van onverwachte gebeurtenissen en is niet gebonden aan dezelfde beperkingen om gegevens te verzamelen die gelden voor gerandomiseerde onderzoeken met een controlegroep. Hierdoor is PR een pragmatische stap op weg naar de slimme operatiekamer, met op wetenschappelijk bewijs gestoelde ondersteuning van de chirurg en zijn team. Hieronder worden de in dit proefschrift voorgestelde technieken beschreven, ingedeeld naar fase (pre-, intra- en post-operatief).

Om de pre-operatieve planning van operaties te verbeteren is het van belang van tevoren de complexiteit (en daarmee de verwachte duur, waarschijnlijke complicaties en benodigde middelen) van de ingreep te kunnen inschatten. Ook kan deze inschatting de chirurg helpen bij het kiezen van de aanpak, bijvoorbeeld door voorafgaand aan de operatie extra voorzorgsmaatregelen te treffen of de patiënt door te verwijzen naar een gespecialiseerde chirurg of ziekenhuis indien een zeer complexe operatie wordt verwacht. We laten zien dat op basis van reeds van tevoren beschikbare gegevens deze complexiteit kan worden ingeschat. Tijdens dit onderzoek zijn classificatiesystemen getraind met en hierna gebruikt voor reeds verzamelde gegevens van patiënten die een laparoscopische cholecystectomie (LAPCHOL) hebben ondergaan. Hierbij kan de complexiteit met een nauwkeurigheid van 83% worden bepaald met een LDC of SVM classificatiesysteem. Ook de vier belangrijkste typen gegevens zijn hierbij geïdentificeerd, namelijk: aanwezig zijn van een ontsteking, dikte van de galblaaswand, geslacht en BMI score.

Om de veiligheid en efficiëntie intra-operatief te verbeteren, zijn ziekenhuizen op zoek naar automatische registratiesystemen die de workflow in de operatiekamer vastleggen. In dit proefschrift stellen we een techniek op basis van gegevens over het gebruik van instrumenten voor. Op basis van deze gegevens worden vervolgens de chirurgische stappen en fase en hun duur afgeleid. Door dit toe te passen in een realtime systeem kan het chirugisch team realtime worden ondersteund en kan de planning voor de volgende patient realtime worden bijgestuurd.

Het gebruik van instrumenten kan op allerlei manieren worden gevolgd, onder andere via specifieke sensoren voor het detecteren van video, beweging en RFID tags. Sensoren die kunnen worden toegepast in de operatiekamer moeten aan een veeltal eisen voldoen, met name omtrent sterilisatie, en mogen de chirurg en zijn team niet belemmeren in het uitvoeren van de operatie. We stellen een systeem voor dat het gebruik van instrumenten detecteert op basis van endoscopische video. De instrumenten worden in de video realtime herkend aan de hand van weefsel compatibele en steriliseerbare kleurmarkeringen. Dit systeem kan zowel een enkel als meerdere instrumenten tegelijkertijd volgen en heeft als resultaat een logfile. Hierin is geregistreerd wanneer en hoe lang ieder instrument is gebruikt.

De logfiles van meerdere operaties kunnen vervolgens worden gebruikt voor workflow mining en het detecteren van onverwachte gebeurtenissen. We laten

zien dat de consensus workflow kan worden afgeleid van meerdere logfiles met multiple sequence alignment. Hierbij voldoet de afgeleide consensus workflow aan de hoofdstappen voor een LAPCHOL, zoals beschreven in bestaande best practices. Middels global pair-wise alignment kunnen afwijkingen van een operatie ten opzichte van de consensus workflow worden gevonden. Deze afwijkingen zijn normaliter kleine variaties in de uitvoering van de chirurgische ingreep, maar kunnen ook het resultaat zijn van ernstige complicaties of fouten.

Een mogelijkheid om de post-operatieve efficiëntie te verbeteren is het vooraf betrouwbaar kunnen inschatten van de duur van het verblijf van de patiënt in de verkoeverkamer. Hiervoor kunnen bestaande peri-operatieve gegevens worden gebruikt, waaronder demografische kenmerken van de patiënt (o.a. leeftijd en geslacht), gegevens verzameld tijdens te operatie (o.a. duur en complicaties), toegediende medicijnen, patiënt comorbiditeit en behandelende chirurg. Lineaire regressie is gebruikt samen met least squares en least absolute shrinkage and selection operator (LASSO) regressie. Middels forward feature selection methoden zijn vervolgens de meest bepalende typen gegevens bepaald om de duur van het verblijf in de verkoeverkamer te voorspellen. We laten zien dat door gebruik te maken van deze gegevens bij de voorspelling deze 12 tot 18 minuten nauwkeuriger kan worden gedaan ten opzichte van de gemiddelde waarde. Door deze voorspelling te verbeteren middels realtime gegevens, voornamelijk door het meten van de zuurstof verzadiging, kan deze voorspelling tijdens het verblijf in de verkoeverkamer steeds betrouwbaarder worden. Dit is een onderwerp van toekomstig onderzoek.

In dit proefschrift zijn er verschillende toepassingen van patroonherkenning technieken voorgesteld en toegepast om chirurgische workflow te registreren, analyseren en voorspellen. Het hier getoonde werk heeft geen directe bijdrage geleverd aan het verminderen van fouten en verbeteren van patiëntveiligheid in en rondom de operatiekamer. Echter de technieken die hier zijn ontwikkeld kunnen wel worden gebruikt om de standaardisatie van chirurgische workflow te ondersteunen en zo zowel fouten te verminderen als de efficiëntie te verbeteren. Het is daarnaast mogelijk deze technieken - met beperkte aanpassingen - in andere medische toepassingsgebieden te gebruiken.

Loubna Bouarfa

Acknowledgements

The completion of a PhD thesis is a lot more than running experiments and writing articles, it is a journey of personal growth, interaction and development. One of the joys of completion is to look over this journey and remember all colleagues, friends and family who have helped and supported me along this long and now fulfilled journey of four years. I would like to thank the following people in particular.

Jenny, thank you for giving me the opportunity to do this PhD. You did not only gave me the support and supervision that a graduate student can expect from her professor, but you gave me trust and freedom that a researcher needs to cross the threshold from supervised work to independent work. Although your time is scarce, your office door is always open for everyone, you can always make time for any question or concern. I also much enjoyed being with you in international conferences and meeting your broad international network. Jenny, together we achieved very good results and I'm looking forward to our continued collaboration in the future.

Pieter, thank you for being my second promotor. I much enjoyed working with you and the colleagues from the robotic vision Lab: Oytun, Maja, Berek, Xin, Boris and Wouter. I learned lot about Robotics from my interaction with your team. I am grateful for giving me the opportunity to participate in teaching the course "probabilistic robotics", I experienced teaching as a fun experience and love to teach more in the future.

My internship in Munich was an exciting experience. I am most grateful for Prof.Dr. Nassir Navab for giving me the opportunity to do this internship and for introducing me to the CAMP lab, the MITI group and to Professor Heinz Lemke. Nassir, it was a pleasure to work with you and with your team and colleagues in Munich and I'm looking for more collaboration in the future. Thanks goes also to Prof.Dr. Heinz Lemke for the nice discussions about medical workflow.

For my internship in Munich, my special thanks go to the MITI group: Prof.Dr. Hubertus Feussner, thank you for welcoming me into the MITI family. I am so

impressed by your accurate performance as a surgeon. You are one of the few surgeons that admires and demonstrates the value of surgical workflow research; your performance in the OR demonstrates that surgery is a profession of accuracy. Dr. Armin Schneider, I am very grateful for your great hospitality. You welcomed me in Munich, you prepared a desk for me, you shared data with me, and provided me with all the facilities to make of my stay in Munich the most fruitful three months of my PhD. It was a pleasure to collaborate with you. My thanks to all the other MITI colleagues: Adam, Salman and Sarah.

My internship in Trondheim was another exiting adventure. Midway the winter of 2011, I spend one month at SINTEF. Special thanks to Andreas Seim and Thomas Longo for inviting me and arranging my stay in Trondheim. It was a pleasure to collaborate with both the process management group and the medical technology group of SINTEF. I'm also very grateful for Andreas for connecting me with the colleagues at the Vanderbilt medical centre in Nashville. For the colleagues at the Vanderbilt medical centre, Special thanks to Dr. Jesse Ehrenfeld, and Dr. Brian Brothman; thank you for sharing your huge dataset with me, it was a pleasure to work with you and to learn how hospital workflow on the other side of the Atlantic works.

A special word of gratitude for the user- and doctoral- committee of this PhD project: Special thanks to (Prof.) Dr. Laurents Stassen, thank you for the observation studies you allowed me to run in your OR and for sharing your laparoscopic video data with me, many thanks also to all the Reinier de Graaf team for welcoming me into their OR. Many thanks to the other members of the use-committee: Prof.Dr. F.W. Janssen, Prof.Dr.ir. C.A. Grimbergen, Dr. B.C. Verdouw, and Dr. H.C.P.M. van der Valk. Also a special thanks to all the other doctoral committee members: Prof.dr. M.A. Neerincx, Dr. T.Weijters and Prof.dr.J.Klein, thank you for agreeing to serve as a doctoral committee member for my PhD. defence.

There are three fellow researchers from TU Delft whom I would specifically like to thank for their contributions in their thesis: Dr. David Tax for the thorough discussions and fruitful collaboration in two papers. David you are my example of the "good" researcher, I have much respect for you as a researcher and as a colleague. Oytun Akman, besides being a colleague, you are a close friend, I am very proud of your career achievements and I have had the pleasure to work and collaborate with you during my PhD. I hope we will continue our friendship from wherever we are. Dr. Dimitra Dodou thank you for reviewing my papers and for the thorough discussions about the future academic perspectives. Dimitra, I like your passion for multidisciplinary research, and hope we stay in touch.

In four years some colleagues have become more than just colleagues. I'm grateful to our secretary team: Diones, Anouk and Dineke, you give the department a cheerful atmosphere, help us with every matter, never forget any opportunity to celebrate and share cakes in the 'kantoortuin'. Also I like to thank these colleagues in particular: Maja, Magda and Berek, we were more than just collegues the previous four years, we enjoyed good moments in and outside the

office hours. I hope to stay in touch with you. Special thanks to Paul Breedveld and Amir Zadpoor for your comments on my thesis and Paul for your nice ideas about the cover. I also extend my gratitude to all other office colleagues in TU Delft.

Moving towards more personal acknowledgements, I would like to execute many thanks towards all my family [1] and my friends [2] for their help, love, friendship and patience. Special thanks goes to my mother-in-law; Ingrid, I'm very grateful for you babysitting Jad on Wednesdays and for the couple overnights when we go abroad. Special thanks to my young brothers Soufiane and Marouane, with your love and our special bond, I never felt alone or lonely, even when writing my PhD thesis.

I am particularly indebted to my parents, my husband Jeroen and my son Jad for their unwavering support and encouragement on all fronts. My dear parents, from you I learned how to dream, and that every dream is just a matter of time. Neither time nor distance have prevented you from been there for me when I needed you, and without you none of this would have been even possible. My dear Jeroen, the past four years would not have been so wonderful and remarkable without you in my life. Thank you for all your love, patience and support. Without your support this thesis would not have been finished in time. In this four years, we already achieved lot together, and I'm looking forward to our coming achievements. My dear little Jad, you were born during my PhD, and made me see things in perspective. You are the joy of my life.

Loubna Bouarfa
Delft, April 2012

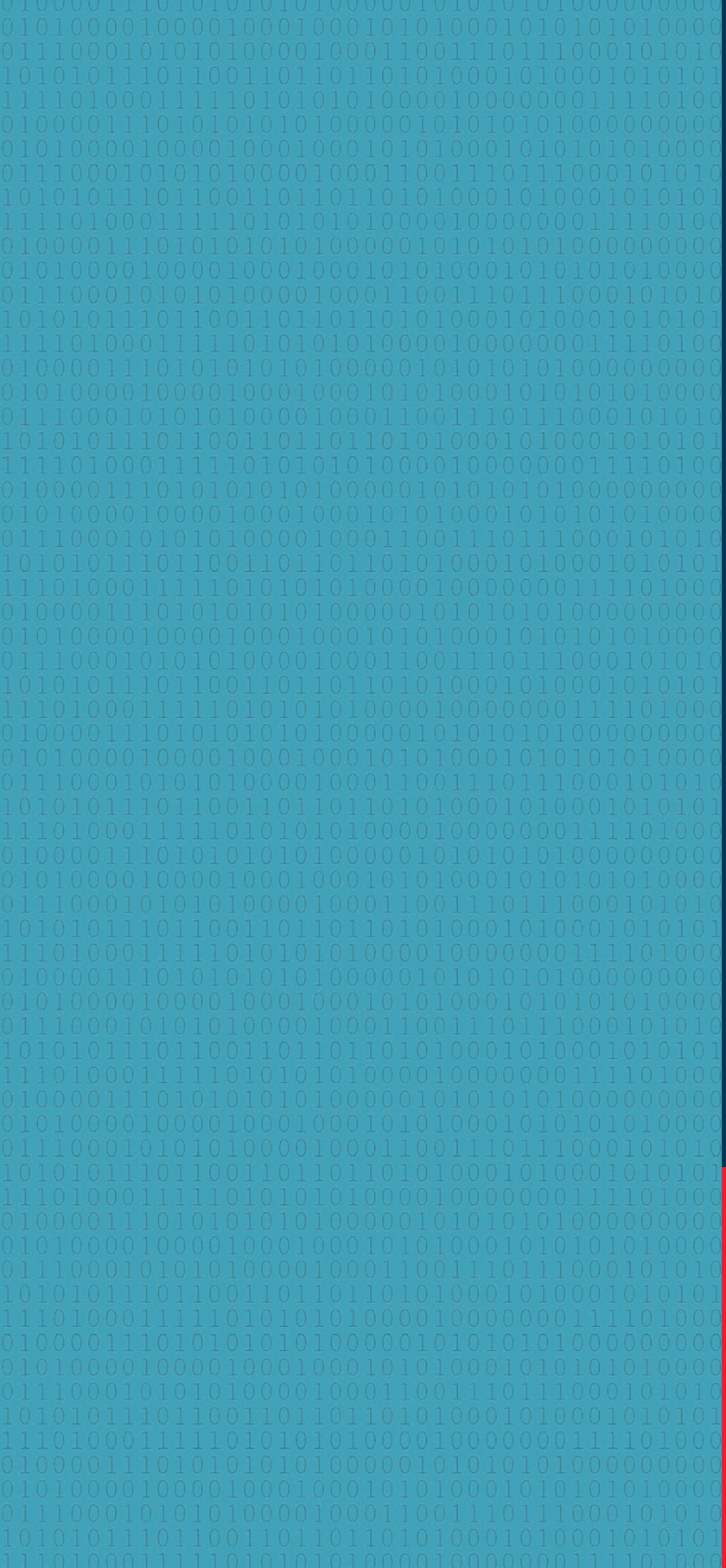
[1] You know who you are.
[2] You also know who you are.

Curriculum Vitae

Loubna Bouarfa was born in Meknès, Morocco on June 26, 1983. She obtained her baccalaurèat de science experimental from Omar Ibn Elkhatab secondary school in Meknès, Morocco, in 2000. In the same year, she was admitted to the electrical engineering department at Delft University of Technology (TU Delft) in The Netherlands. After finishing the language preparatory course, she starts the bachelor program at the electrical engineering department in September 2001. In October 2007 she obtained her Master of Science degree in Media Knowledge Engineering from TU Delft. Her graduation project, at the Information & Communication Theory Group, was about video fingerprinting, under the supervision of Prof.dr.ir. Inald Lagendijk.

In November 2007, Loubna joined the Biomechanical Engineering Laboratory of Delft University of Technology as a PhD student, under the supervision of Prof.dr. Jenny Dankelman and Prof.dr.ir.Pieter Jonker. Her research work spans the area of surgical workflow and data analysis using pattern recognition tools. During this PhD, she did an internship at the Technical University of Munich within two research groups: The Computer Aided Medical Procedures group and the Minimally Invasive Interdisciplinary Therapeutically Intervention Laboratory. She was also a visiting researcher for one month at the department of Technology and Innovation at SINTEF, Trondheim, Norway. During the last year of her PhD, she collaborated with the Perioperative Data Systems Research group at Vanderbilt Medical Centre, Nashville, USA.

In June 2012, Loubna will join the Hamlyn Center at Imperial college London as a research associate within the pervasive computing group. Her work will focus on data processing for pervasive healthcare, wearable sensors, well-being and sports applications.



ISBN: 978-94-6169-251-1