A METHODOLOGY FOR QUANTITATIVE PERFORMANCE EVALUATION IN MINIMALLY INVASIVE SURGERY

by

PAUL BRADLEY MCBETH

B.Sc., The University of Calgary, 1999

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Department of Mechanical Engineering)

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

February 2002

© Paul Bradley McBeth, 2002
In presenting this thesis in partial fulfilment of the requirements for an advanced
degree at the University of British Columbia, I agree that the Library shall make it
freely available for reference and study. I further agree that permission for extensive
copying of this thesis for scholarly purposes may be granted by the head of my
department or by his or her representatives. It is understood that copying or
publication of this thesis for financial gain shall not be allowed without my written
permission.

Department of MECHANICAL ENGINEERING

The University of British Columbia
Vancouver, Canada

Date MARCH 6, 2002
Abstract

The objective of this work is to establish a methodology for reliably quantifying the performance of an expert surgeon in laparoscopy with the long-term goal of validating surgical simulations. A validated simulation will allow us to quantitatively assess surgeon performance and evaluate new tool designs. Quantitative performance and skill assessments are critical for evaluating the progress of surgical residents and the efficacy of different training programs. Current evaluation methods are subjective and potentially unreliable, so there is a need for objective methods to evaluate surgical performance. We identify a feasible method to measure kinematic and postural data in the live operating room setting. We used an optoelectronic motion analysis system to acquire postural data and tool tip trajectories of one expert surgeon over a period of four months. To assess reliability of performance measures, we created a hierarchical decomposition diagram describing the procedure in terms of surgical tasks, tool sequences and fundamental tool actions. Using tool tip kinematic data and postural data we extracted characteristic measures of performance and compared these measured distributions using the Kolmogorov-Smirnov statistic. For the most part our performance measures (with the exception of kinematic measures) show consistent reliability over time by a trained surgeon and little effect from patient variability, and so are likely reliable measures of performance. An expanded set of reliable kinematic measures will form the basis for quantifying surgical skill and should be useful in validating surgical simulations for use in training, certifying surgeons and designing and evaluating new surgical tools.
Table of Contents

Abstract ..................................................................................................................................................... ii
Table of Contents ......................................................................................................................................... iii
List of Tables ................................................................................................................................................ vii
List of Figures ............................................................................................................................................... viii
Acknowledgements ..................................................................................................................................... xi
Chapter 1 - Introduction and Literature Review ....................................................................................... 1
1.1 Introduction and Objectives .................................................................................................................. 1
1.2 Overview of Minimally Invasive Surgery (MIS) ..................................................................................... 2
1.3 Limitations in Minimally Invasive Surgery ............................................................................................ 4
1.4 Performance Evaluation ......................................................................................................................... 5
1.5 Training and Simulators ......................................................................................................................... 6
1.6 Simulator Validation ............................................................................................................................... 8
1.7 Ergonomics ............................................................................................................................................ 9
1.8 Analysis of Performance Measures ....................................................................................................... 10
1.9 Kinematic Assessment Tools ............................................................................................................... 11
1.10 Research Questions ............................................................................................................................. 12
Chapter 2 - Quantitative Methodology of Evaluating Surgeon Performance in Laparoscopic Surgery .......... 14
2.1 Introduction .......................................................................................................................................... 14
2.2 Methods ............................................................................................................................................ 15
   2.2.1 Framework .................................................................................................................................. 15
      2.2.1.1 Project Methodology ........................................................................................................... 15
      2.2.1.2 Hierarchical Decomposition ................................................................................................. 20
      2.2.1.3 Performance Measures ........................................................................................................ 21
         2.2.1.3.1 Time ............................................................................................................................... 22
         2.2.1.3.2 Kinematics ...................................................................................................................... 22
         2.2.1.3.3 Ergonomic / Postural .................................................................................................... 23
         2.2.1.3.4 Event Sequencing .......................................................................................................... 24
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.1.4 Difference Measures</td>
<td>24</td>
</tr>
<tr>
<td>2.2.1.4.1 Kolmogorov-Smirnov Statistic</td>
<td>24</td>
</tr>
<tr>
<td>2.2.1.4.2 Sequencing Delta Measure</td>
<td>26</td>
</tr>
<tr>
<td>2.2.1.5 Evaluating Reliability</td>
<td>27</td>
</tr>
<tr>
<td>2.2.1.5.1 Procedural Specific Variability</td>
<td>28</td>
</tr>
<tr>
<td>2.2.1.5.2 Context Comparisons</td>
<td>30</td>
</tr>
<tr>
<td>2.2.1.6 Clinical Interpretation</td>
<td>31</td>
</tr>
<tr>
<td>2.2.2 Experimental Protocol and Data Acquisition and Processing</td>
<td>32</td>
</tr>
<tr>
<td>2.2.2.1 Equipment</td>
<td>32</td>
</tr>
<tr>
<td>2.2.2.2 Protocol</td>
<td>34</td>
</tr>
<tr>
<td>2.2.2.3 Missing Data and Signal Processing</td>
<td>35</td>
</tr>
<tr>
<td>2.2.2.4 Joint Angle Processing</td>
<td>36</td>
</tr>
<tr>
<td>2.2.3 Reliability Analysis – Clipping Task</td>
<td>37</td>
</tr>
<tr>
<td>2.2.4 Context Comparison – Clipping Task</td>
<td>37</td>
</tr>
<tr>
<td>2.2.5 Clinical Interpretation – Clipping Task</td>
<td>38</td>
</tr>
<tr>
<td>2.3 Results</td>
<td>38</td>
</tr>
<tr>
<td>2.3.1 Feasibility</td>
<td>38</td>
</tr>
<tr>
<td>2.3.2 Reliability</td>
<td>40</td>
</tr>
<tr>
<td>2.3.3 Context Comparisons</td>
<td>46</td>
</tr>
<tr>
<td>2.3.4 Clinical Interpretation</td>
<td>48</td>
</tr>
<tr>
<td>2.4 Discussion</td>
<td>48</td>
</tr>
<tr>
<td>2.4.1 Data Acquisition</td>
<td>49</td>
</tr>
<tr>
<td>2.4.2 Hierarchical Decomposition</td>
<td>49</td>
</tr>
<tr>
<td>2.4.3 Reliability Analysis</td>
<td>50</td>
</tr>
<tr>
<td>2.4.3.1 Procedural Specific Variability</td>
<td>50</td>
</tr>
<tr>
<td>2.4.3.2 Context Comparisons</td>
<td>52</td>
</tr>
<tr>
<td>2.4.4 Clinical Utility</td>
<td>53</td>
</tr>
<tr>
<td>2.4.5 Future Work</td>
<td>55</td>
</tr>
<tr>
<td>2.5 Conclusion</td>
<td>56</td>
</tr>
<tr>
<td>Chapter 3 - Repeatability of Biomechanical Stresses in Laparoscopic Surgery</td>
<td>57</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>57</td>
</tr>
<tr>
<td>3.2 Methods</td>
<td>58</td>
</tr>
<tr>
<td>3.2.1 Equipment</td>
<td>58</td>
</tr>
<tr>
<td>3.2.2 Kinematic Arm Model, Calibration and Joint Angle Calculations</td>
<td>58</td>
</tr>
<tr>
<td>3.2.3 Comparative Analysis (KS Statistic)</td>
<td>59</td>
</tr>
<tr>
<td>3.2.4 Ergonomic Stress Scoring System</td>
<td>60</td>
</tr>
<tr>
<td>3.2.4.1 Modified Discrete RULA Method</td>
<td>62</td>
</tr>
<tr>
<td>3.2.4.2 Modified Continuous RULA Method</td>
<td>62</td>
</tr>
<tr>
<td>3.2.5 Post-Operative Calculations</td>
<td>64</td>
</tr>
<tr>
<td>3.3 Results</td>
<td>64</td>
</tr>
</tbody>
</table>
List of Tables

Table 2.1 – Summary of data available from each procedure ........................................... 39
Table 3.1 – General comments regarding each procedure .................................................. 71
Table 3.2 – The average time taken for the major surgical tasks ....................................... 74
List of Figures

Figure 1.1 - The use of reliable measures of surgical performance .............................................. 2
Figure 1.2 - Illustration of minimally invasive surgery ................................................................. 3
Figure 1.3 - Reduced DOF of motion of the MIS tool tip .............................................................. 4
Figure 1.4 - Illustrations of commonly used surgical simulators ................................................... 7
Figure 1.5 - The Kolmogorov-Smirnov statistic ............................................................................. 10
Figure 2.1 - Typical uses of the comparative analysis methodology ............................................... 16
Figure 2.2 – A demonstration of our validation assessment technique ........................................... 19
Figure 2.3 - Laparoscopic cholecystectomy hierarchical decomposition ........................................ 21
Figure 2.4 – The tool tip and x,y,z frame of the MDMArray ......................................................... 23
Figure 2.5 - The Kolmogorov-Smirnov statistic ............................................................................. 25
Figure 2.6 – Example of a state transition diagram and probability transition matrix. .... 26
Figure 2.7 – Performance measure reliability analysis ................................................................. 29
Figure 2.8 – Global performance evaluation measures for a clipping task .................................... 31
Figure 2.9 - Data collection components ....................................................................................... 33
Figure 2.10 - Model of the surgeon’s arm. ...................................................................................... 34
Figure 2.11 - Graphical representation of missing data interpolation ............................................. 40
Figure 2.12 – Stage transition diagram for the subtask sequence ................................................ 41
Figure 2.13 (a) – Completion time reliability analysis ................................................................. 43
Figure 2.13 (b) – Completion time reliability analysis ................................................................. 43
Figure 2.14 (a) – Joint angle distribution reliability analysis ......................................................... 44
Figure 2.14 (b) – Joint angle distribution reliability analysis ......................................................... 44
Figure 2.15 (a) – Similarity sequence of clipping subtasks ............................................................ 45
Figure 2.15 (b) – Similarity sequence of clipping subtasks ............................................................ 45
Figure 2.16 (a) – Kolmogorov-Smirnov difference statistic of completion time performance measures ........................................................................................................... 46

Figure 2.16 (b) – Kolmogorov-Smirnov difference statistic of kinematic performance measures................................................................................................................................. 47

Figure 2.16 (c) – Kolmogorov-Smirnov difference statistic of mean membership action variables. ................................................................................................................................. 47

Figure 2.17 – Performance measure reduction for a clipping task ...................................... 48

Figure 2.18 – Example of measuring the performance of a new surgeon by comparing performance in a single procedure to a reference database from our expert surgeon…… 54

Figure 2.19 – Resolution of completion time ........................................................................ 55

Figure 3.1 - The Kolmogorov-Smirnov statistic ................................................................... 60

Figure 3.2 – R.U.L.A. Posture zones. Zero reference positions are also shown. ................. 61

Figure 3.3 – Discrete and continuous modified RULA posture scores. ............................... 63

Figure 3.4 - The values labeled on each marker represent the percentage of time it was tracked over all seven procedures after interpolation.................................................. 65

Figure 3.5 - Graphical representation of cumulative probability distributions of each joint and for each procedure................................................................. 66

Figure 3.6 - Graphical representation of the cumulative probability distributions of wrist ulnar / radial deviation. ................................................................. 67

Figure 3.7 – Variations in neutral posture measurements found experimentally in the lab and from the OR data................................................................. 68

Figure 3.8 - Graphical representation of the cumulative probability distributions of elbow flexion and extension ................................................................................. 69

Figure 3.9 - Graphical representation of normalized cumulative probability distributions of each joint and for each procedure ................................................................. 70

Figure 3.10(a) - Graphical representation of cumulative probability distributions correlating procedures to groupings of joint angle distributions ............................ 71

Figure 3.10(b) - Graphical representation of cumulative probability distributions showing types of variability from procedure to procedure........................................... 72

Figure 3.11 – Benchmarking the difference measures using the K-S statistic on a difference scale ......................................................................................................................... 73
Figure 3.12 – Cystic duct dissection (CDD) stress levels for each joint angle normalized and averaged over seven procedures, including the continuous and discrete NWPSS. ... 74

Figure 3.13 – Gallbladder dissection (GBD) stress levels for each joint angle normalized and averaged over seven procedures, including the continuous and discrete NWPSS. ... 75

Figure 3.14 – Gallbladder removal (GBR) stress levels for each joint angle normalized and averaged over seven procedures, including the continuous and discrete NWPSS. ... 75

Figure 3.15 – Shifted continuous and discrete NWPSS averaged over seven procedures for each joint in each surgical phase. ................................................................. 78
Acknowledgements

I would like to thank all of those who have supported me throughout the course of this project. I am grateful to my supervisor, Dr. Antony Hodgson for his knowledge, insight and enthusiasm. I am particularly grateful to Dr. Alex Nagy, for sharing with me his wealth of knowledge in the area of laparoscopy and for putting up with relentless OR experiments and seemingly painful joint calibrations. I would like to thank Dr. Karim Qayumi for is assistant in helping analyze data and providing clinical interpretation.

Many thanks to Jeff Stanley and others from Northern Digital Inc. (Waterloo, ON) for their assistance with Polaris and Mona Charman, Betty Barns, and all the nurses and staff at Vancouver hospital for their cooperation and interest in my project.

I would like to thank all my colleagues of the Neuromotor Control Laboratory. I am indebted to my many of them and especially grateful to Cam Shute, Scott Illsley, Chris Plaskos, and Willem Astma. Cam, thank you for reminding me of my priorities and keeping things in perspective, those days of skiing and climbing always seemed to bring clarity. Scott, thank you for getting up early on those rainy winter mornings to help with my OR experiments. I could not have done it without you. Chris, thank you for reminding me of the subtleties of life and what academia is all about. Willem, thank you for all those enlightening philosophical discussions on just about everything.

Many thanks to all of my friends of whom I bounced off ideas or to whom I may have whined to on occasion, and to everyone else who helped me along the way.

Most of all, I would like to thank my family, and especially my parents, for their support and absolute confidence in me. This thesis is dedicated to them, without whom none of this would have been even possible.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).
Chapter 1
Introduction and Literature Review

1.1 Introduction and Objectives

Over the past decade minimally invasive surgery (MIS) has become an important part of modern medicine. New surgical techniques and protocols and developments in surgical equipment have allowed many conventionally open procedures to be done using MIS techniques. Although MIS has been widely adopted over the past decade many of these new procedures are more stressful for the surgeons. These procedures generally take longer and require surgeons to use awkward postures, and the tools limit surgeons’ dexterity and range of motion.

Surgical simulators are playing an increasingly important role in training and credentialing surgeons and have a potentially significant role to play in designing and evaluating new surgical equipment. As simulators come to serve a more crucial role in the early training of a surgeon, it is important to ensure that the skills they learn are directly relevant to the practice of surgery in the operating room (OR). This is important because of the comparative ease of making objective measurements in a simulator; performance there is used as a significant component in evaluating or certifying surgeons. Also, if we can show that tools are used essentially in the same way in the simulator as in the OR then we can improve and test new tool designs before introducing them into the OR.

The long-term goal of this research is to establish a methodology for quantifying the validity of surgical simulators for the purpose of evaluating new tool designs and assessing skill and performance of surgeons (see Figure 1.1). We would like to show that time spent practicing motor skills in surgical simulators translates into quantifiable improvements in surgical skill in the OR and that performance in the simulator is a reliable measure of surgical skill. Furthermore, we would like to show quantifiable differences between different tool designs. To do this we propose to identify corresponding measures of performance in the OR and surgical simulators.
2. Equipment Evaluation and Design

3. Performance Measures

Simulator Validation
Surgeon Certification

Figure 1.1 - The use of reliable measures of surgical performance in simulator validation, surgeon certification and equipment evaluation and design.

The focus of the research presented in this document is to identify ways of measuring performance of an expert surgeon during laparoscopic cholecystectomies (gallbladder removal). The goal is to identify reliable measures of performance and to benchmark these measures in a non-stereotyped complex motor task performed by an expert surgeon in the OR. The benchmarked performance measures will be useful for establishing the level of performance of an 'expert surgeon' and identifying performance measures which are reliable enough to use in simulator validation and evaluating new surgical equipment.

1.2 Overview of Minimally Invasive Surgery (MIS)

The concept of minimally invasive surgery (MIS) is not new; in fact, the idea has been around since the beginning of the 20-century (Nagy 1992). However, it wasn’t until the development of fiber optic light sources in the 1960’s that the potential for laparoscopy was realized. Further developments in camera technology and tool design has led to widespread laparoscopic procedures since the late 1980’s. Since that time, MIS has advanced into an important part of modern general surgery. In the United States 60-80% of all abdominal surgical procedures are now conducted using MIS techniques (Taylor 1995). Figure 1.2 shows a typical operating room setup for laparoscopic procedures.
The development of MIS techniques has resulted in significant benefits for patients. For example, in comparison with open techniques laparoscopy offers the patient (1) minimal scarring, (2) reduced hospital stays, (3) reduced postoperative pain, and (4) faster recovery (Treat 1996, Perissat 1995). Patients are often discharged from the hospital one or two days after a procedure and complication rates for laparoscopic cholecystectomy are low, ranging from 2% to 11% (Catalano 1997).

Surgeons, on the other hand, must deal with limitations in their ability to perform relatively simple motor tasks. Commonly used laparoscopic tools limit surgeons’ dexterity and range of motion, thus forcing them to use awkward postures to carry out relatively simple tasks (Person 2001). Laparoscopic instruments are considered by many surgeons to be difficult to use and “non-user-friendly” (Treat 1996). These problems, coupled with increased operative times (approximately 30% larger than standard open procedures (Treat 1996)), have caused the use of laparoscopy to be restricted primarily to comparatively simple procedures such as cholecystectomies and hernia repairs.
1.3 Limitations in Minimally Invasive Surgery

In laparoscopy, the reduction in the number of degrees of freedom of the surgical tool limits the surgeon's dexterity and range of motion (Tendick 1995). A laparoscopic tool has only 4 DOF as compared to a 6 DOF open tool. The motion of the tip is constrained to pitch, yaw, roll, and plunge as shown in Figure 1.3 (Person 2000). Not only does the surgeon experience motion reversal with the tools but they also experience changes in the location of the fulcrum as the tool is moved in and out of the abdominal cavity (Treat 1996). Surgeons have an indirect view of the surgical site via a camera and a 2D video display. The lack of depth perception places limits on difficult maneuvers such as suturing and knot tying. Tactile feedback is significantly reduced in laparoscopic surgery, which often leaves the surgeon guessing at various aspects of the surgery. These drawbacks dissuade surgeons from undertaking more complex procedures using MIS techniques and so have prompted research to improve surgeon performance and efficiency during MIS procedures.

![Figure 1.3 - Reduced DOF of motion of the MIS tool tip. DOF are roll, pitch, and yaw about the fulcrum created by the entry portal and plunge through the port. (Source: Tendick 1995).](image)

Laparoscopic surgery is considered technology-dependent and costly (Straface 1995). Limitations in video system technology and the dexterity of the tools cause difficulties in laparoscopy (Hunter 1997, Pellegrini 1997, Way 1995). New technologies are being developed to address limitations of conventional laparoscopic surgery including; improved vision systems (Tendick 1995), multifunctional and more dexterous

1.4 Performance Evaluation

A means of assessing performance and skill in MIS procedures is potentially useful for surgeon trainee evaluation, simulator validation, and surgical tool evaluation. Current evaluation methods are subjective and potentially unreliable so there is a need for objective and standardized methods to evaluate surgical performance (Rosser 1998, Winckel 1994, Lentz 2002, Chung 1998).

Structured skills assessment forms represent the gold standard for surgical evaluation. These forms allow comprehensive intraoperative performance measurement, which address both the planning level and psychomotor skills of a surgeon. Winckel used this method to show the difference between junior and senior residents (Winckel 1994). Eubanks used a similar procedure to evaluate videotapes of a procedure (Eubanks 1999). However, patient variability and the stressful environment of the OR make it difficult to quantify the level of technical surgical skill based solely on these evaluation methods. It is also difficult to identify the level of a surgeon's technical surgical skills because these are not explicitly quantified during a structured assessment using the same kinds of measurements typically made in simulators (e.g., tasks completion time, precision measures, etc.).

The efficiency of a procedure can be evaluated in terms of speed and quality. Some aspects of performance are easy to quantify while others are difficult. The time required to perform a procedure is easy to measure, and has been used by several research groups in an attempt to evaluate surgical performance (Chung 1998, Derossis 1998, Fried 1999, Hanna 1998, Hodgson 1999, Keyser 2000, Rosser 1997, Starkes 1998, Taffinder 1999).
A number of measures of quality have been suggested, some of which are more easily measured than others. Fried investigated surgeon performance in both simulated tasks and animal models (Fried 1999). His assessment is based on a penalty scoring system where scores are assigned for various errors executed during the test such as tool tip deviations, presence of bleeding, and excess suturing. Joices' team investigated surgical performance using a Human Reliability Analysis to identify inter- and intrastep errors. These are errors made at the procedural and motor skill levels. These types of analyses are useful and can co-exist with our techniques.

One valuable approach for performance evaluation was recently demonstrated on a porcine model and is based on an analysis of tool-tip force/torque signatures (Rosen 2001). The study uses a Markov modeling technique in conjunction with a structured classification of tool actions to classify and assess surgeon performance. They showed that they could correctly classify surgeons into two experience levels based on the similarity of the Markov models representing their low-level tool-tissue interactions to models derived from reference groups representing the two experience levels. Rosen identifies other feedback data such as tool position as likely alternative or supplementary bases of performance evaluation.

1.5 Training and Simulators

Success in laparoscopy is highly dependent on the surgeon's proficiency and experience (Perissat 1995). Apprenticeship training is the most common method of providing experience and improving proficiency of new surgeons. Unfortunately, the surgical trainer cannot control which patients enter the hospital, potentially limiting a trainee's exposure to a range of surgical problems.

Training on animal models may avoid some of the stresses and time constraints of apprenticeship training, although the anatomy often differs from that of humans, one cannot often replicate a disease state in the animal, and maintaining an animal care facility is expensive. There are also moral and ethical concerns related to training on live
animals. Some places, for example the United Kingdom, have banned the use of animals for surgical training (Lirici 1997).

Surgical simulators are becoming more widely used as they can represent a wide range of physical models and virtual reality computer simulations. Recently, virtual reality systems with haptic interfaces have been developed. Simulators are designed to emphasize either or both of psychomotor skills (e.g., suturing drills) and the more cognitive, process-related aspects of surgery (e.g., the CyberPatient process simulator program). Figure 1.4 shows examples of simulators commonly used.

*Figure 1.4 - Illustrations of commonly used surgical simulators.* Computer based virtual reality systems (left), anatomical models (middle), and bench-top simulators (right). (Source: www.laparoscopy.com)

Surgical training programs are moving towards a more structured curriculum in which surgical models and simulators play a more important role both in developing surgical skill and in evaluating and certifying trainees. To increase the acceptance and integration of simulators into training programs, it is important to show that time spent on simulators can actually replace O.R. experience. For the evaluation and certification of trainees on simulators it is important to show a performance evaluation based on a simulated task strongly correlates with the appropriate aspects of live surgical performance. It is crucial to ensure that the lessons learned on the simulator and the measurements of technical skill made there are truly relevant and applicable to performance in the operating room.
1.6 Simulator Validation

A validated surgical simulation is important for (1) surgical skill training, assessment and certification, and (2) equipment evaluation and design. The challenge in validating a surgical simulation is to guarantee that performance results on the simulator represent performance in the OR. For the purpose of surgical skill training, assessment and certification, the value of a simulation depends on how closely it captures the elements of perceptual-motor and cognitive skills used in an actual task (Sanders 1991). To validate a simulation, we must ensure that a surgeon treats the simulation in as many relevant and measurable respects as possible, the same way they treat a live patient.

A number of research groups are attempting to validate surgical simulators primarily for the purpose of training and evaluating new surgeons. Most groups select some mechanical manipulation tasks, which they consider comparable to real surgical tasks. Completion time is the most common measure of performance (Chung 1998, Derossis 1998, Fried 1999, Hanna 1998, Hodgson 1999, Rosser 1997, Starkes 1998, Taffinder 1999). Some groups have used error and precision measures as methods to quantitatively assess performance (Derossis 1998, Fried 1999, Hanna 1998, Keyser 2000, Taffinder 1999). It has typically been shown that more experienced surgeons score better on the trainers than novice surgeons (O'Toole 1998). Comparatively little work has been done to quantitatively show the relationship between the performance on simulators and performance in the OR.

Current methods for establishing correspondence between simulated and actual OR tasks are poor. The authors tend to agree that the chosen simulated tasks are relevant to OR tasks and that performance in simulated and real tasks are related; this notion that there should be substantial similarity between simulations and real situations is known as "face validity" (Reznick 1993) and is typically considered the first and least rigorous of about four validation steps (least rigorous because it is typically not quantified, but simply emerges from a consensus of application experts). It is my contention, however, that by using motion capture instrumentation for measuring the details of how a surgeon moves, we can produce a quantitative similarity measure between simulated and real tasks – the first quantitative measure of face validity.
The research reported here uses detailed quantitative measurements of surgical tool motions and surgeon posture from both live operative procedures and surgical simulations, broken down by surgical context and motor task classification, to establish correspondences between the simulation and the operative tasks. The goal is to develop simple desktop simulations (perhaps based on existing ones – e.g., Derossis 1998) which simulate tool and arm motions similar to those surgeons use in actual surgery. Such detailed kinematic comparisons cannot be done using manual measurement techniques and so will require the kind of intraoperative measurement technique we have described here.

1.7 Ergonomics

In addition to skill evaluation, postural assessment is important in evaluating improved instrumentation and validating simulators. Postural assessments are common in repetitive task workplace settings, but only a small number of studies have been done in the operating room. Radermacher et al. was one of the first to assess minimally invasive surgeries and suggested improvements for workplace design. Berguer showed that the postures in laparoscopic surgery are more stressful than in conventional open surgery using video and force plate measurement techniques (Berguer 1997). In these past studies, visual or video based observations have formed the bases of postural assessment. Recently, Person conducted a pilot study to investigate the feasibility of using an optoelectronic system to measure joint angle position and assess ergonomic stress during a live OR procedure (Person 2001). From his study it remains unclear as to how strongly ergonomic stress is affected by procedural and patient specific factors.

Surgeon injuries resulting from poor instrument design are also common (Horgan 1997). Several musculoskeletal disorders have been associated with laparoscopic surgery including: thumb pain ("Laparoscopic surgeons’ thumb") (Neuhaus 1997), hand and elbow pain, lower back pain, and shoulder pain (Buschbacher 1994). Surgeon fatigue, digital neuropathies, and finger discomfort are common conditions among laparoscopic surgeons (Mueller 1993). Research to improve surgeon comfort has resulted in improved handle designs. Different handle designs are available including: (1) axial handles with
trigger rings, (2) inline handles with trigger rings, and (3) pistol grip handles with a flat shank or trigger rings (Matern 1999). Many of these new tool designs have been shown to benefit surgeons by reducing fatigue and numbness in hands and arms (Hasson 1993, Neuhaus 1997, Horgans 1997).

1.8 Analysis of Performance Measures

The goal of this research is to establish reliable measures of performance which can be used to validate surgical simulators, assess performance of surgeons, and quantify performance of new surgical instrumentation. As discussed in Section 1.4 there are a variety of measures which can be used to assess performance. In this research we investigate time, kinematic, joint angle, and sequencing performance measures.

To compare performance data across two settings (i.e. OR vs. simulator, or Tool A vs. Tool B in a simulator or the OR) we use a hierarchical decomposition technique similar to Cao (1996) to decompose the entire operation into phases, stages, tasks, subtasks, and actions. We extract performance measures from individual tasks and subtasks, and assign a measure of dissimilarity to the data sets obtained from the two settings. The Kolmogorov-Smirnov (K-S) statistic D is used to quantify the discrepancy between the two distributions of the measured performance variable. The K-S statistic indicates the maximum vertical difference between two cumulative distribution functions (see Figure 1.5).

![Figure 1.5 - The Kolmogorov-Smirnov statistic, D, is defined to be the maximum vertical difference between two cumulative distribution functions. The cumulative distribution function for a finite data set is represented as a staircase.](image-url)
The Kolmogorov-Smirnov statistic is useful because it is nondimensional, makes no \textit{a priori} assumptions about the nature of the statistical distribution of the data, and is also comparatively insensitive to the presence of outliers in the data. The K-S statistic is often used to test the null hypothesis that two sets of data were drawn from the same underlying distribution. Since our two data sets are obtained from two separate tasks we would expect the two distributions to be different. For our analysis we are interested in the measure of the difference between the two distributions. Significant differences will be interpreted as indicators of where we need to refine our simulation to better match the corresponding operating room (OR) task, to quantify differences in performance levels and to evaluate new tool designs.

1.9 Kinematic Assessment Tools

We have chosen a complementary approach to work done by Rosen by investigation of performance measures using kinematic data (Rosen 2001). Our approach is to investigate the kinematic measures related to tool position and upper limb position. To track surgical tools and limbs in space, we require a high frequency tracking system. Several commercially available systems are available for tracking the position and orientation of markers in space. Such systems include magnetic, ultrasonic and optoelectronic trackers, and accelerometer/gyroscopic base systems. In addition, goniometers can be used to produce joint angle estimates. Each system has its own unique benefits and drawbacks.

Magnetic systems provide continuous, high frequency sampling (60Hz) however, each tracker requires a fixed wire attachment to the tool interface system. In addition, magnetic systems are strongly influenced by external ferrous materials and electromagnetic fields, which are often present in the OR. Ultrasonic systems are continuous high frequency wireless systems which utilize high frequency sound waves to track objects. Ultrasonic trackers are also faced with the line-of-sight requirement of the transmitter and receivers. This system is also influenced by ambient noise sources. Optoelectronic systems provide wireless high frequency data collection and are easily sterilized. Hybrid systems are capable of tracking both passive (wireless) and active
(IRED's) markers. These systems are limited by its line-of-sight restriction and interference from external infrared light sources. Accelerometer/gyro based systems also provide high frequency sampling, although they also require fixed wire attachments and often require extensive calibration.

1.10 Research Questions

The long term goal of this research is to quantify the fidelity of surgical simulators with applications to surgical training, surgical skill evaluation, equipment evaluation and equipment design for minimally invasive surgery. The objective of the portion of this research reported in this thesis is to establish a methodology to quantitatively assess the performance of an expert surgeon for a laparoscopic cholecystectomy procedure. A number of questions are addressed in this research:

1. What is a feasible way to quantitatively measure surgeon performance? What measures of surgeon performance can be used to reliably quantify performance in a non-stereotyped OR task? What are the characteristic performance measures which describe the performance of an expert surgeon and how can these measures be presented in a clinically useful way?

2. What is the repeatability of biomechanical stresses on an expert surgeon performing a laparoscopic cholecystectomy using a quantitative motion analysis technique? Is a motion analysis based measure of stress strongly affected by procedural and patient specific factors? How can we quantitatively assess the performance of a surgeon with new surgical instrumentation?

Chapter 2 presents a feasible method to measure kinematic data in the live operating room setting. An assessment of the effectiveness of an optoelectronic motion analysis system in obtaining kinematic data in the OR is provided. In addition we address the effect of interpolated and extrapolated data on measurement accuracy. This chapter also presents research in the quantitative evaluation of surgeon performance. We investigated
the similarity of kinematic performance measures based on a hierarchical decomposition of surgical tasks in a laparoscopic cholecystectomy. In addition event sequencing is used to quantify the repeatability of progressing through the logical steps of a procedure.

Chapter 3 presents the repeatability of biomechanical stresses on an expert surgeon performing a laparoscopic cholecystectomy. The effect of patient specific factors on measurements of biomechanical stress is investigated and tested to see if a single procedure provides a reliable estimate of stress experienced by a surgeon.

Chapter 4 summarizes the findings of the two main sections of the thesis and proposes future studies using the methodology developed here to validate surgical simulations.
Chapter 2
Quantitative Methodology of Evaluating Surgeon Performance in Laparoscopic Surgery

2.1 Introduction

Over the past decade, minimally invasive surgery has become an integral part of modern medicine. However, developments in teaching and evaluation methods have not kept pace. Quantitative performance and skill assessments are critical for evaluating the progress of surgical residents and the efficacy of different training programs. They are also important for validating surgical simulations and in assessing new surgical tool sets and techniques. Current evaluation methods are subjective and potentially unreliable, so there is a need for objective methods to evaluate surgical performance (Chung 1998, Rosser 1998, Winckel 1994).

Surgical simulators are playing an increasingly important role in training and objective evaluation of surgeons. Commonly used surgical simulators include bench top trainers, animal models and virtual reality systems. A wide selection of performance measures have been suggested in the literature to quantify performance including: completion time, frequency of errors, force/torque signatures, low level event sequencing, and tool tip positioning (Chung 1998, Derossis 1998, Fried 1999, Hanna 1998, Hodgson 1999, Rosser 1997, Starkes 1998, Taffinder 1999, Torkington 2001). One valuable approach for performance evaluation was recently demonstrated on a porcine model and is based on an analysis of tool tip force/torque signatures (Rosen 2001). The study uses a Markov modeling technique in conjunction with a structured classification of tool actions to classify and assess surgeon performance. Rosen identifies other feedback data such as tool position as a basis of performance evaluation. As simulators come to serve a more crucial role in the early training of a surgeon, it is important to ensure that the skills they learn are directly relevant to the practice of surgery in the operating room (OR). This is important because of the comparative ease in making objective measurements in a simulator.
The long-term objective of this research is to validate surgical simulations by showing that time spent practicing psychomotor skills in simulators relates to improvements in surgical skill in the live operating room setting, and that measured performance in the simulator is a reliable measure of surgical skill. In general, current techniques for showing validation demonstrate that experienced surgeons perform better than novices in simulated tasks and that practice leads to improved simulator performance (Rosser 1997, Starkes 1998, Chung 1998, Derossis 1998, Hanna 1998, Fried 1999, O’Toole 1999). In some cases these assessment methods have shown that simulators can be used to evaluate progress (Starkes 1998, Chung 1998) and even more compelling, that practice in the simulator alone leads to improved performance in a pig model (Fried 1999). This implies that carefully chosen simulated tasks can be used to replace OR time for learning psychomotor skills. Our goal is to show quantitatively the performance level of any particular surgeon based on measurements taken in a simulated setting.

This chapter outlines a methodology for evaluating surgeon performance and establishes a framework for a quantitative analysis of the performance of an expert surgeon performing a laparoscopic cholecystectomy (gallbladder removal). The objective is to identify a feasible method to measure kinematic data of a surgical tool and postural data in the live operating room setting and to assess the reliability of this analysis method based on a hierarchical decomposition of surgical tasks. We investigate the effect of patient and procedural variability on our assessment of performance and show examples of how this method can be used to validate surgical simulators.

2.2 Methods

2.2.1 Framework

2.2.1.1 Project Methodology

The goal of this research is to develop a methodology for validating surgical simulations. The methodology developed here is intended as a tool to measure and highlight differences of surgical performance in different settings. Primarily, we are interested in using this method for the purpose of validating surgical simulations. However, we have
also identified its utility for clinical certification and equipment evaluation and design. Consider the following example: to help validate a surgical simulation we may look at the performance of an expert surgeon in the live OR and compare this to his or her performance in a surgical simulation. If we can show that performance for all relevant aspects in a simulation is essentially the same as performance in the OR, then we have a validated simulation or face validity (Allen, 1986). Finding discrepancies between two settings can help us identify and correct problems with the simulation design, help an instructor to correct a trainee’s technique, or help tool designers make improvements to instrumentation.

To demonstrate our comparative analysis approach, consider the following examples where this method might be applied (See Figure 2.1):

Figure 2.1 - Typical uses of the comparative analysis methodology for simulator validation, certification, and Equipment evaluation & design. The segments circled represents contributions of this thesis.
Simulator Validation: For the purpose of validating surgical simulations we look for discrepancies in selected measures of performance from the OR and the simulator. Having identified these discrepancies, we can go back and make adjustments to the simulator and conduct a new evaluation. The process is continued until we are satisfied the simulator matches the relevant aspects of the OR procedure we are interested in.

Surgeon Certification: Accurately quantifying surgeon performance is important to training and certification of trainees and surgeons. Using our methodology, we can extract performance measures of a trainee in the OR (or simulator) and compare his or her performance to a reference database of performance measures from trainees at various levels in their training programs. This will allow us to quantitatively assess the skill level and identify problematic areas where improvements are required. For example, if we look at a particular surgical task, such as applying a clip, we may notice a surgeon spending more time doing a particular aspect of the clip application. Using our methodology, we can highlight specific problematic areas and suggest ways of making improvements.

Instrumentation Design and Evaluation: Using a validated simulation, tool designers can evaluate the performance of new instrumentation and be confident their performance will transfer to the OR. As in the validation process of surgical simulators, the evaluation of new instrumentation is an iterative process where the performance of the new tool is compared with an existing database of performance measures for past tool designs.

For the purpose of making these comparisons, we establish an organizational framework describing a surgical procedure. Surgery in general can be thought of as collection of tasks organized in a particular sequence, but diversions may occur due to unexpected or variable events or conditions, such as bleeding. A hierarchical decomposition allows us to study difference in tasks even when the procedures are not exactly the same (Cao, 1996). These differences are found in the low level details of tool movements which are common to all procedures. The hierarchical decomposition is useful as an organizational
framework for making various context comparisons. For example, simulators are typically designed to model only specific tasks rather than the entire procedure. The decomposition approach will allow us to make comparisons between surgical tasks and simulations of those tasks.

The basis of our methodology for comparative analysis requires quantifying face validity. To quantify face validity we select meaningful measures of performance and compare their distributions across settings. See Figure 2.2 (top). A useful performance measure will demonstrate reliability and incorporate details of surgical performance which are important to the quality and efficiency of a procedure. Establishing the reliability of a particular performance measure will depend on its behavior in different settings (comparatively unaffected by variations between patients and procedural differences and yet sensitive to real differences such as training effects and improved instrumentation) and its subjective acceptance by clinicians. Under ideal conditions we would expect a reliable performance measure to behave essentially the same in settings, which are identical in all relevant aspects. For quantifying the difference between performance measure distributions across various settings we use the Kolmogorov-Smirnov (K-S) statistic (Press 1992). The K-S statistic provides a difference measure between two distributions, which produces values from 0 (similar) to 1 (different). See Figure 2.2 (bottom). The K-S statistics was chosen because it considers the entire distribution rather than other statistics which only test for differences in mean (t-test) or variance (f-test) (Von Mises, 1964).
By developing a reference database of comparative analyses (OR vs. simulator, surgeon A vs. surgeon B, etc) we establish meaning or context to the difference scale. For example, if we evaluated an expert surgeon from day to day on a surgical simulation we would expect the distribution of performance measures to be extremely similar since the simulator has not changed and the expert surgeon is fully trained. On the other hand if we evaluated a novice surgeon we would expect some degree of difference attributable to learning effects. If we evaluated an expert surgeon only in the OR and compared performance measure distributions from procedure to procedure we would expect to see differences attributable to patient and procedural variability. These measures of difference provide meaning to the difference scale and thus establish reference points for clinical interpretation.

Finally, we are concerned with our ability to distinguish differences among our performance measures. In particular this is important when trying to resolve differences between surgeons of difference skill in the OR or simulator. In the OR, our measurements are influenced by procedural and patient variability. As a result we want to determine the resolution of our measurements so we can reliably detect differences between performance measures.
The contributions of this methodology presented in this chapter include: (1) a method to reliably measure completion time, kinematic, event sequencing, and joint angle performance measures, (2) establishing the reliability of selected performance measures based on the evaluation of an expert surgeon during a live OR procedure and showing the influence the procedural specific factors on variability, and (3) providing examples of context comparisons to illustrate the utility of this method.

To demonstrate the application of our proposed methodology we selected an example task (applying a clip), which is performed comparatively frequently (on average seven times per procedure) and compared the characteristic measures associated with its component actions across the clipping segments recorded.

### 2.2.1.2 Hierarchical Decomposition

Many of our chosen performance measures are specific to only certain segments of the procedure. For example, the distance for a tool tip relocation maneuver can only be defined for low-level movement segments. To provide an organizational structure for our data analysis, we created a hierarchical decomposition describing the procedure in terms of surgical phases and stages, tool tasks and subtasks, and fundamental tool actions, with each level containing more specific details than the previous (see Figure 2.3). This technique is based on a decomposition approach originally described by Cao, with our system having been modified to improve generality and incorporate additional kinematic features on low-level tool movements (Cao, 1996). The decomposition design was based on the standard elements of laparoscopic cholecystectomies and was developed in consultation with two expert surgeons. The decomposition has five levels; the phase level (1) outlines the global goals of the procedure. The stage level (2) outlines local goals required to complete each phase. The task level (3) involves the use of a single tool. The subtask level (4) describes how the surgical tool is moving inside the patient. Finally, actions (5) describe kinematic features (or potentially force/torque signatures) of stereotyped fundamental movements such as reaching or sweeping. This five-level hierarchical decomposition provides the foundation for a quantitative analysis of surgeon
performance since it allows us to quantify standard aspects of procedures, thereby reducing variability in assessing a procedure as a whole. This framework is useful for making comparisons between simulators and the OR as many simulators only model small segments of the entire procedure (i.e., Derossis's mechanical manipulation tasks). Operational definitions such as start and end points are presented in Appendix D.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Subtasks</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trocar Prep</td>
<td>Push</td>
<td>- Push</td>
</tr>
<tr>
<td></td>
<td>- Spread</td>
<td>- Spread</td>
</tr>
<tr>
<td></td>
<td>- Hold</td>
<td>- Hold</td>
</tr>
<tr>
<td>Cystic Duct</td>
<td>Translate</td>
<td>- Translate</td>
</tr>
<tr>
<td>Dissection</td>
<td>Grasp</td>
<td>- Grasp</td>
</tr>
<tr>
<td>Galbladder</td>
<td>Release</td>
<td>- Release</td>
</tr>
<tr>
<td>Dissection</td>
<td>Retract</td>
<td>- Retract</td>
</tr>
<tr>
<td>Galbladder</td>
<td>- Pull</td>
<td>- Pull</td>
</tr>
<tr>
<td>Removal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closure</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2.3** - Laparoscopic Cholecystectomy Hierarchical Decomposition. The decomposition has five levels, each containing more specific details than the previous.

### 2.2.1.3 Performance Measures

There exists a wide range of performance measures for evaluating surgeon performance. In consultation with two expert surgeons we have identified completion times, tool tip kinematics, joint angles, and event sequencing as potentially relevant measures of performance. Other measures such as force/torque signatures and frequency of errors, as reported in the literature, can easily be integrated into our methodology. The following sections outline our selected performance measures used to assess performance.
2.2.1.3.1 Time

The most commonly used performance measures in past studies have been task completion time (Chung 1998, Derossis 1998, Fried 1999, Hanna 1998, Hodgson 1999, Keyser 2000, Rosser 1997, Starkes 1998, Taffinder 1999). A time-based measure alone, however, will not necessarily capture all aspects of motor performance. As a result we look at additional performance measures which we feel capture importance aspects of motor performance. The hierarchical decomposition is used to define start and stop points at each level as outlined in Appendix D.

2.2.1.3.2 Kinematics

Using kinematic measures to assess surgeon performance is relatively new. Torkington’s group uses electromagnetic trackers to measure distance, number of movements, and speed for surgeon’s hand movements in simulated tasks and Rosen suggested its importance for evaluating skill (Torkington 2001, Rosen 2001). In our approach we use tool tip kinematic data to extract: travel distance, mean, minimum, and peak tool tip velocity, minimum and peak acceleration, jerk cost (Nelson, 1986), and RMS straight-line deviations.

In addition to the kinematic performance measures we investigated the composition of tool tasks and subtasks in terms of the 12 basic and fundamental tool actions. In total there are 72 feasible combinations of these 12 actions. Figure 2.4 outlines the kinematic details of selected states. We described each of the 12 basic actions using fuzzy membership functions based on instantaneous tool tip position and velocity. For the purpose of establishing performance measures we use the mean action membership values of tasks and subtasks in each of 12 prototypical actions. The mean action membership value was chosen because of the difficulty in automatically detecting segments of action states. In Cao’s work verbal kinematic descriptions of tool actions are found based on video data (Cao, 1996). The analysis of video data is an extremely
laborious and time consuming process which lead to our approach of using action membership functions. Appendix E outlines details of calculating membership values.

![Diagram of tool tip and x.y.z frame calibration]

**Figure 2.4** – The tool tip and x.y.z frame is calibrated to the MDMArray on the tool handle. The tool tip coordinate frame is used to define kinematic details of each state.

### 2.2.1.3.3 Ergonomic / Postural

Postural assessments in minimally invasive surgery are rare, despite widespread acknowledgement that strain injuries do occur to surgeons (Radermacher 1996, Berguer 1997, Person 2001). Training surgeons to use the correct posture is important for long-term occupational health. We evaluate posture to ensure subjects use the same stance and arm positioning in simulators as they would in the OR. A postural evaluation will allow us to compare joint angles for selective tasks and make adjustments to the position of the simulator to match the OR. Based on upper limb modeling by (Person 2001) we calculate the following joint angles: shoulder abduction/adduction, flexion/extension, and internal/external rotation, elbow flexion/extension, pronation/supination and wrist flexion/extension and radial/ulnar deviation of the dominant arm.
2.2.1.3.4 Event Sequencing

Finally, we investigated high-level (procedural) surgical event sequences as measures of performance. Our approach to event sequencing modeling differs from the low-level Markov Modeling demonstrated by Rosen (Rosen, 2001). A Markov Model is a random model describing the probability of moving from one state to another based on the transition probabilities and previously visited states at regularly sampled intervals. The Markov Modeling technique explicitly represents time, which is not necessarily important for describing the sequencing of surgical events. Our approach is to separate time and sequencing into individual performance measures. This provides greater applicability for identifying relevant surgical sequencing. In contrast to Rosen’s approach of using low-level details of random force/torque signatures, we propose an alternative method based on transition probabilities of structured (non-random) event sequences. This measure operates on the level of the surgical process and allows us to identify variations in the course of a surgical procedure.

2.2.1.4 Difference Measures

2.2.1.4.1 Kolmogorov-Smirnov Statistic

Our methodology relies on extracting distributions of performance measures from different contexts. For any particular variable of interest we use the Kolmogorov-Smirnov (K-S) statistic \( D \) to quantify the discrepancy between two distributions of the measured variable. The K-S \( (D) \) statistic represents the maximum vertical difference between cumulative probability distributions (see Figure 2.5); this measure ranges from 0 (similar) to 1 (different). The associated \( p \) value expresses the probability that the two measured distributions arise from the same underlying distribution and provides an indication to the significance of any differences found in the two distributions.
The K-S statistic is useful because it is non-dimensional, makes no \textit{a priori} assumptions about the statistical distributions of the data, and is comparatively insensitive to the presence of outliers (Hodgson, 2002). The K-S statistic is commonly used to test the null hypothesis that two sets of data were drawn from the same underlying distribution. In our case we use the K-S statistic as a measure of the difference between the two distributions and use this measure of difference to quantify differences in performance measures.

Our data is based on two discrete sets of measures; thus, the $D$ value we find is an estimate of the difference between the two sets. In order to assign a confidence interval to our estimate of $D$ we use a bootstrapping method based on the single set of measurements. This method involves finding the distribution of $D$ values by randomly resampling the dataset and applying the K-S statistic $n$ times where $n \gg 0$. Bootstrapping is a way to test the reliability of our dataset (Efron, 1986).

Due to time and cost restraints we often have a finite collection of performance measures. Analysis of the K-S statistic suggests when the two finite sets of measures are small or when the difference in means is low, our estimated $D$ value is consistently greater than the actual value (Hodgson, 2002). Our goal is to compensate for size of data set to enable direct comparison across $D$ measures. Using numerical experimentation, the following function gives an approximation of $\bar{D}$ (our estimated $D$ value):
\[
\overline{D} = D + \overline{D}_0(N) - \frac{0.7742D + 0.0078}{1 - D(1 - 1/\overline{D}_0(N))}
\]  

[2.1]

Where: \( \ln \overline{D}_0(N) = k_1 + k_2 \ln N \)

\( k_1 = 0.0261, k_2 = -0.471. \)

Given \( \overline{D} \) (our estimate), we can solve Equation 2.1 for \( D \), which is adjusted for \( N \) where \( N \) is the number of elements in the data vector. This approximation for \( D \) is based on comparing equally sized distributions and normally distributed data.

2.2.1.4.2 Sequencing Delta Measure

Transition state diagrams were used to represent these event sequences and describe the probability of transitions between different states (Hennie, 1968). These diagrams provide an indication as to how often various transitions occur during an event sequence (see Figure 2.6). The transition probabilities are calculated by dividing the number of transitions between one state and a second state by the total number of transitions from the first state to all other possible states. The transition probabilities of the state diagram are reduced to a Transition Probability Matrix \((TPM)\). Each cell of the transition probability matrix \((TPM_i)\) represents the transition probability from one event to another (event \( i \) to event \( j \)).

![State Transition Diagram and Probability Transition Matrix](image)

**Figure 2.6** – Example of a state transition diagram and probability transition matrix.
We developed a sequencing discrepancy measure to quantify differences between event sequences. For any particular event sequence (ie. sequence of subtasks during a clipping task) we use a sequence difference measure $\delta$ to quantify the discrepancy between the two event sequences. For example, consider comparing a single procedure event sequence ($P_i$) to a reference event sequence ($P_{\text{ref}}$) to establish a measure of difference $\delta$. The definition of the sequencing discrepancy measure $\delta$ is given by:

$$\delta = \frac{\sum_{i=1}^{n} (P_{\text{ref},i} - P_{\text{sample},i})^2 w_i}{2\sum_{i=1}^{n} w_i},$$

where $P_{\text{ref},i}$ and $P_{\text{sample},i}$ are the $i^{th}$ transition probabilities of $TPM_{\text{ref}}$ and $TPM_{\text{sample}}$ respectively, $w_i$ represents the frequency of the $i^{th}$ event in the reference sequence and $n$ is the total number of transition probabilities.

The sequence difference measure $\delta$ is analogous to our K-S ($D$) statistic as it is non-dimensional and varies on a scale of zero to one (similar to dissimilar). In addition, our measure satisfies the definition of a metric (ie. for sequences $S$, $T$, and $U$: $\delta(S,T)>0$, non-negative; $\delta(S,T)=0 \iff S=T$, coincidence; $\delta(S,T)=\delta(T,S)$, symmetry; $\delta(S,T)+\delta(T,U)\geq\delta(S,U)$, triangle inequality) (Mannila, 1998). The sequence difference measure is based on two finite sequences resulting in an estimate of the difference between the two. We use a bootstrapping approach to assign a confidence interval to the estimate.

2.2.1.5 Evaluating Reliability

To demonstrate the reliability of the proposed analysis procedure, we must show the performance measures are comparatively unaffected by variations between patients and procedural differences and yet sensitive to real differences such as training effects and improved instrumentation. A database of performance measure data from various skill-
leveled surgeons in the OR and simulator is required for establishing overall reliability. In this document we (1) quantify the effects of procedural and patient specific variability for a single surgeon and (2) use contextual comparisons to show sensitivity in an attempt to establish the reliability of our proposed performance measures for an expert surgeon.

2.2.1.5.1 Procedural Specific Variability

To demonstrate the effect of procedural specific variability we use the K-S statistic to measure the difference between a distributions of $D$ values related to a single procedural comparison (with patient variability) to our best estimate of the distribution of $D$ values without procedural specific variability. Any difference between these distributions is an estimate of procedural variability.

To illustrate this analysis technique, consider total completion time for a clipping task. Each procedure produces a distribution of clipping completion times represented as a cumulative probability distribution ($\text{CPD}_i$). The underlying distribution of completion times is established by grouping together clipping task completion times of all procedures to get a total cumulative probability distribution ($\text{CPD}_T$) as shown in Figure 2.7 (left). The $\text{CPD}_i$'s of each procedure are compared with the $\text{CPD}_T$ ($\text{CPD}$ for all procedures excluding $\text{CPD}_T$) to produce a distribution of measured $D$ values ($\text{CDF}(D_{\text{measured}})$). The $\text{CDF}(D_{\text{measured}})$ represents the range of $D$ values we would expect to get if we measured the same surgeon in the OR on an arbitrary patient. See dotted line shown in Figure 2.7 (right). To establish our best estimate of a $D$ distribution, where we would expect no procedural specific contributors to variability, the $\text{CPD}_T$ is randomly resampled in lengths equivalent to the number of occurrences in each procedure (for example, a clip task occurs on average seven times per procedure thus the resample length is seven) and compared against itself without replacement for $r$ resamples where $r \gg 0$. This produces a distribution of reference $D$ values ($\text{CDF}(D_{\text{reference}})$) which is the range of $D$ values we would expect to get if we measured this surgeon on a “standard” (identical) patient with no specific contributors to variability. See solid line shown in Figure 2.7 (right).
Figure 2.7 – Performance measure reliability analysis – Cumulative Probability Distribution's (CPD) of total clipping time for individual procedures (CPD_i) and an accumulated grouping (CPD_T) (left). CPD of $D_{reference}$ and $D_{measured}$ for increasing number of procedures analyzed (right). The calculation process for evaluating the $D_{reference}$ and $D_{measured}$ distributions (bottom).

The shift between $D_{reference}$ and $D_{measured}$ is an indication of the effect of patient variability. This provides a measure of our ability to detect differences with the same surgeon and different procedures. The K-S statistic is used to measure the difference between $D_{reference}$ and $D_{measured}$ ($D(D_{reference}, D_{measured})$) and check if procedural and patient specific contributors are significant. Significance is established using both analytical and
bootstrapping techniques outlined in Appendix E. In combination with the K-S statistic we also measured the horizontal shift \((D_{\text{horz}})\) to quantify the magnitude of procedural specific variability. \(D_{\text{horz}}\) is measured by subtracting the means of each distribution of \(D\)'s.

To further investigate our ability to resolve differences and to identify effects of procedural specific variability we applied the above analysis technique to multiple procedures. In this case \(CDF(D_{\text{measured}})\) is generated by comparing \(\text{CPD}_{j2c}\) to \(\text{CPD}_{\text{TOC}}\) where \(\text{CPD}_{j2c}\) represents all two procedure combinations of \(\text{CPD}_j\) and \(\text{CPD}_{\text{TOC}}\) represents all procedures excluding \(\text{CPD}_{j2c}\). The \(CDF(D_{\text{measured}})\) is calculated by resampling in lengths equivalent to the number of occurrences in two procedures and compared against itself without replacement. Figure 2.7 (right) illustrates the effect of multiple procedure comparisons.

A technique analogous to this method is used to evaluate the procedural specific variability of event sequencing performance measures. This modified technique replaces the \(D\) distributions \((CDF(D_{\text{measured}})\) and \(CDF(D_{\text{reference}})\) with \(\delta\) distributions \((CDF(\delta_{\text{measured}})\) and \(CDF(\delta_{\text{reference}})\)\). All other steps are the same. A demonstration of this technique is presented in Appendix E.

**2.2.1.5.2 Context Comparisons**

To demonstrate reliability of our selected performance measures (in our case the application of the K-S analysis technique) we assess differences by comparing distributions of performance measures obtained from multiple executions of a particular task or subtask in two different contexts. These contextual comparisons allow us to highlight differences which may, for example, show the procedural variability which is useful for establishing meaning to our difference scale.
2.2.1.6 Clinical Interpretation

For the purpose of surgical education it is useful to describe overall performance rather than individual performance measure evaluations because our analysis produces so many of them. In this section we provide a method of reducing our performance measure collection to an overall evaluation parameter. Figure 2.8 shows a conceptual illustration of this proposed analysis technique. Assessing a surgeon’s skill level requires comparing their collection performance measures to a reference database, resulting in a collection of $D$-values.

![Figure 2.8 - Global performance evaluation measures for a clipping task. A collection of $D$ and $\delta$ values is presented for each performance measure comparison. A weighted average based on performance measure reliability is used to reduce each performance measure of a single performance value for each performance measure type (time, kinematics, posture, sequencing, etc...). Each performance measure type is plotted in a multidimensional performance space](image-url)
To start, we must show reliability of our performance measures in distinguishing differences in various settings such as the OR and simulators for both experts and novices. From these analyses we extract ranked ordered lists of performance measures based on their reliability. For example, consider our collection of kinematic performance measures (tool tip travel distance, peak velocity, etc...). These measures are ranked according to their ability to distinguish performance differences, their influence from patient variability and subjective clinical importance (i.e. clinically smoothness may be more important than tool path distance). The ranked ordered lists are established for each performance measure type (i.e. kinematics and posture) within each subtask, resulting in four performance measures (time, kinematics, posture, and sequencing). For each subtask within a task, the performance measures are summarized according to clinical importance. The summary is based either on maximum values (for the purpose of highlighting areas requiring improvement) or a weighted average. The weighting of each subtask is related to its importance for overall task completion and its potential for risk of injury. For example, there may be more risk to injury in the tissue manipulation, and clip application subtasks than the approach and withdrawal, thus giving a higher weighting. The summarized measures are plotted in a multidimensional performance space. These values are averaged to give an overall evaluation parameter.

This method is useful for comparing performance of different skilled surgeons in both the OR and simulators. It is beneficial for highlighting areas requiring improvements thus allowing surgical educators to correct a trainee’s technique by suggesting specific areas needing improvement.

2.2.2 Experimental Protocol and Data Acquisition and Processing

2.2.2.1 Equipment

An optoelectronic motion analysis system was used to acquire postural data and tool tip trajectories at frequencies of ~20 Hz. For this study, we used a Northern Digital Polaris Hybrid Optical Tracking System capable of tracking the 3D position of both active infrared light emitting diodes (IRED’s) and passive retro-reflective markers with an accuracy
of ~0.2-0.3 mm. The optoelectronic system was chosen over other motion capture systems because it is well tolerated by surgeons, accurate, and parts in the OR field are easily sterilizable. The Polaris system was connected to a standard PC (800 MHz AMD Duron) running custom-designed data acquisition software written in Matlab.

In addition to the optoelectronic system, we recorded video images of the surgery using both a laparoscope and an external camera focused on the surgeon. The images from these two sources were time stamped and recorded onto standard VHS tape; from these images, we could later determine the stage of the operation and correlate it with the detailed motion measurements. The laparoscope system was comprised of a standard 10 mm - 0° surgical laparoscope, camera and illuminator (Stryker Endoscopy). These components are arranged and connected in the OR as illustrated in Figure 2.9. All equipment used in the operating room was tested and approved by Vancouver Hospital’s Biomedical Engineering Department and sterilized with ethylene oxide when appropriate.

**Figure 2.9** - Data collection components – Polaris Hybrid Optical Tracking System and Video Recording Equipment (left). Custom designed data acquisition software written in Matlab (right).

The surgical instrumentation used for the study consisted of a standard set of reusable laparoscopic tools designed for cholecystectomy (gall bladder removal) procedures. The same tool set was used in each procedure. None of the tools were specifically designed to enhance ergonomic comfort.
2.2.2.2 Protocol

One expert surgeon was evaluated in seven clinical laparoscopic cholecystectomies over a period of four months at Vancouver Hospital (January - April, 2001). For each procedure, clinical assistants and OR staff varied. There was no a priori selection of the patient group, which included both males and females with varying weights. We attached sterilized marker arrays to the surgeon's torso and dominant arm (on the proximal and distal forearm and on the hand), as shown in Figure 2.10 (left, middle). This set of marker arrays enabled us to track and record the surgeon's joint angles at the shoulder, elbow, and wrist. Arrays were secured to the surgeon using elastic and Velcro® harnesses. The torso marker sling, passive marker arm cuffs, and adhesive attachments to the hands were found to be acceptably immobile relative to the torso, arms, and hands in subjective tests (Person 2000). This set of marker arrays enabled us to track and record the surgeon's joint angles at the shoulder, elbow, and wrist.

Figure 2.10 - Model of the surgeon's arm (left). Marker array mounting (schematic and as mounted in the operating room (middle)). Multi-Directional-Marker-Array (MDMArray) (schematic and as mounted on a laparoscopic surgical tool (right). The values labeled on each marker represent the percentage of time it was visible during manipulation segments.
Custom-designed Multi-Directional Marker Arrays (MDMArray) designed to enhance trackability by increasing the angle at which they could be viewed (relative to planar marker arrays) were attached to each laparoscopic tool handle used by the dominant hand, as shown in Figure 2.10 (right). A planar marker array was attached to the non-dominant hand tool and the laparoscope. The MDMArrays were equipped with a quick release clip to allow easy attachment and removal. Six identical MDMArrays were built and attached to six different laparoscopic tools for the entire procedure (grasper, L-grasper, ratcheted-grasper, clipper, scissors, spatula).

One researcher scrubbed into the surgery and attached the sterilized marker arrays to the surgeon's hands, lower arms and torso and laparoscopic tools. This researcher remained to assist the surgeon with any adjustments to the marker arrays or support cuffs. If the surgeon felt uncomfortable wearing the markers at any time during the surgery, they were free to stop the experiment and remove the markers. Immediately postoperatively, the surgeon adopted a reference posture which defined the neutral or zero point for all angles and then performed a set of standard isolated joint motions to provide data used in estimating the joint locations. The tip position of each surgical tool was found using a sphere-fitting calibration procedure in which the tip was placed in a small depression on a fixed surface while the tool was manipulated about that point.

In addition to our marker tracking study, we used video archives from an additional eight procedures. The archived videos extended back two years and all were of the same expert surgeon. The archived video recordings were used only for extracting completion time and event sequencing performance measures.

### 2.2.2.3 Missing Data and Signal Processing

Optoelectronic systems have inherent line-of-sight limitations causing segments of missing data. To alleviate the problems of missing data we developed custom-designed tool marker arrays with multiple faces visible from several directions (the MDMArrays mentioned above). The non-equally sampled data (~20 Hz) was fitted with a cubic spline
using an optimally found smoothing parameter generated by a Generalized Cross Validation (GCV) technique (Woltring, 1982). Spline coefficients were used to resample position data and find higher order derivatives. This method is described in Appendix C. To determine what order of derivatives could be reliably measured, a frequency spectral analysis was completed using a Fast Fourier Transform (FFT) to show signal power levels of the position signal and higher order derivatives. To establish the feasibility of using an optoelectronic system for tool tip and posture tracking in the operating room, we measured the frequency of marker occlusions and estimated the errors associated with interpolating missing data.

2.2.2.4 Joint Angle Processing

The Polaris system reports the spatial position of the marker arrays in space. Joint angles are calculated from the measured position of the marker arrays located on the surgeon’s hand, forearm, upper arm and torso. We used a simplified seven degree-of-freedom (DOF) rigid-body model of the upper limb as shown in Figure 2.10 (left). The model allows us to represent shoulder abduction/adduction, flexion/extension, and internal/external rotation, elbow flexion/extension, pronation/supination and wrist flexion/extension and radial/ulnar deviation. Using this model, we applied a kinematic calibration method based on circle and sphere fitting techniques to locate the approximate centre of each of the shoulder, elbow, and wrist joints. The method locates a joint axis or centre of rotation by fitting a circle (or sphere) to a cloud of raw data points collected from limb markers and expressed in a reference frame attached to an adjacent limb segment. These techniques are well known in the biomechanics and robotics literature (e.g., Halvorsen 1999). Using the joint centre data and a neutral posture reference we calculated joint angles for each joint at each sample instant. For the purposes of our analysis we use the joint angle distributions for the task level as measures of performance for a clipping task.
2.2.3 Reliability Analysis – Clipping Task

To demonstrate and quantify the effect of procedural specific variability we use the K-S reliability assessment method described in Section 2.2.1.5.1. This evaluation method is applied to completion time distributions for the clipping task and subtasks levels. The method is also used for each joint angle distribution for the clipping task. Due to the limited amount of tool tip kinematic data collected, we were unable to apply this analysis technique to our kinematic performance measures. A technique analogous to the K-S reliability assessment is used to evaluate the procedure specific variability of event sequencing performance measures for the subtask event sequence.

By quantifying the effect of procedural variability we are able to establish the resolution of our measurements (our ability to detect differences in different contexts). We expect that as the number of procedures evaluated increases the effects of patient variability is reduced thus allowing us to resolve smaller differences in comparisons of performance measures.

2.2.4 Context Comparison – Clipping Task

We performed several comparative analyses related to the clip application task for completion time and kinematic performance measures. Most commonly, clips are used in the vessel separation (cystic duct (CD) and cystic artery (CA)) stage of the procedure; however, they are also occasionally used to control bleeding throughout the procedure. We would expect there would be no significant difference between performance measures taken from alternate clip applications, as both patient variability and any learning effects will be equally represented in the two data sets. Since the surgeon is fully trained, we would hope for little variation between the first four and the last three procedures studied; any differences would be primarily due to patient variability and the results would represent the reliability of our analysis procedure. Finally, we prospectively investigated several other divisions of the data set in hopes of demonstrating the
existence of significant and interesting differences; for example, we considered whether there might be differences between clips applied for: vessel separation vs. control of bleeding; separating the cystic duct vs. the cystic artery; first placement vs. subsequent placement on the same vessel.

2.2.5 Clinical Interpretation – Clipping Task

Using the performance measure consolidation method outlined in Section 2.2.1.6 we compare the surgeon's performance in a single procedure to all other procedures where we collected motion capture data. For each procedure the performance measure reduction is based on the completion times for both the subtask and task levels and a comparison of joint angle distributions and event sequencing at the task level. We do not use the kinematic measures due to lack of reliability.

2.3 Results

2.3.1 Feasibility

A certain amount of position data was lost during the course of the surgery due to marker occlusion or internal localization errors. Figure 2.10 (left) illustrates the marker array visibility averaged over the seven procedures for individual markers. Using video analysis to separate manipulation from non-manipulation tasks, we found the majority of missing data could be attributed to marker occlusions during segments not directly related to performing a surgical task (e.g. changing instruments). During tissue manipulation tasks, we had complete joint angle data 80% (SD: 10%) of the time and at least one sample per second 92% (SD: 4%) of the time. The trackability of the dominant hand tool depends on which tool is being used (for the last four procedures we tracked the tool 78% (SD: 12%) of the time, before interpolation, during the clipping task. This average was highly procedural specific. In 12 of the 45 tasks had full data sets and 20 successfully recorded more than 90% of the samples, where as in some procedures we obtained virtually no data). Table 2.1 provides a summary of data available for analysis.
Table 2.1 – Summary of data available from each procedure

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Video Data</th>
<th>Kinematic Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Task Time</td>
<td>Sequencing</td>
</tr>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>X*</td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>A</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>B</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>C</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>D</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>E</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>F</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>G</td>
<td>X*</td>
<td>X</td>
</tr>
<tr>
<td>H</td>
<td>X*</td>
<td>X</td>
</tr>
</tbody>
</table>

* No subtask time available

We used a generalized cross validation (GCV) filtering technique to find an optimal smoothing parameter for fitting a cubic spline to position data. The data was resampled at constant intervals and missing data segments interpolated. The RMS error associated with interpolating across varying gap sizes was calculated and is shown in Figure 2.11. An RMS error of 1mm was chosen as an acceptable error which suggests that a maximum gap size of 0.5 sec (10 samples) can be interpolated. We were able to reduce the amount of missing joint angle data by 20% (SD: 2%) (using our interpolation technique with a maximum 1mm RMS error). Using the same interpolation, we reduced the amount of missing data for the right and left hand tool data by 6% (SD: 2%) and 14% (SD: 3%) respectively. The trackability of the right hand tool after interpolation is 82% (SD: 12%).

An analysis of the signal spectral content found each signal having approximately the same cutoff frequency and noise signal properties. Using the Fast Fourier Transform (FFT) of the position data we found the underlying frequency content ($f_c$) to be ~4 Hz. The accumulated power of the position, velocity, acceleration, and jerk signals suggests each signal is appropriate for our analysis as each signal contains at least 93% of its total power at 4 Hz.
2.3.2 Reliability

Despite our efforts to track continuous tool tip data we found our data to have sizable missing data segments making it difficult to extract useful kinematic performance measures. Due to the limited amount of useful kinematic data we are unable to show reliability of our kinematic measures.

To represent the structure of the procedure, we use transition state diagrams to represent the probabilities of moving from one state to another. The state transition diagram shown in Figure 2.12 shows the transition probabilities of a clipping task where the states represent the subtasks of a clip application. The subtask sequence for a clipping task is well defined as the exiting transition probability’s for the following sequence are all relatively high (greater than 85%): start – approach – tissue manipulation – clip application – withdrawal – finish.
Figure 2.12 – Stage transition diagram for the subtask sequence of a clipping task. Transition probabilities are shown and represent the probability of moving to the next state. FSM = Free Space Movement (Based on clipping subtask sequences from 15 procedures)

Having identified performance measures from the available clipping task data we conduct a comparative analysis to establish measures of reliability and context similarity. Figure 2.13 shows the results of our procedural variability analysis of clipping completion time for the task and subtask levels. We measured the difference between $CPD(D_{\text{reference}})$ and $CPD(D_{\text{measured}})$ by evaluating the K-S statistic (vertical distance) and the horizontal shift in means of $CPD(D_{\text{reference}})$ and $CPD(D_{\text{measured}})$ for multiple procedural evaluations. The results of the bootstrapped analysis of the uncorrected and corrected $D$ value are shown in Figure 2.13(a). We used the K-S statistic to test the null hypothesis that the
CPD(D_{\text{reference}}) and CPD(D_{\text{measured}}) were drawn from the same underlying distribution. Here, the null hypothesis of the $D$-value distributions arriving from the same distribution is rejected for all cases. This result is verified by a bootstrapping approach outlined in Appendix E. These results suggest our performance measure is influenced by patient variability and that the effect of patient variability is reduced with increasing number of procedures considered. The horizontal distribution shifts are evaluated by subtracting the mean of each distribution as illustrated in Figure 2.13(b). The uncorrected $D$-values and differences in mean show a decreasing trend with increasing number of procedures analyzed. With the exception of the clip application and the withdrawal subtasks the remaining levels show a general uniformity along the difference scale. The remaining results presented will deal only with the task level.

Results for the joint angle and sequencing performance measure analyses are illustrated in Figures 2.14 and 2.15. The confidence bounds on the joint angle and sequencing measures decrease with increasing number of procedures analyzed. In all cases the difference in distribution means is decreasing. These results suggest that as the number of procedures considered increases our ability to detect differences in performance measure distributions is improved because the amount of patient variability is reduced.
Figure 2.13 (a) – Completion time reliability analysis for task and subtask level of a clip application. KS (D) value between CPD(D<sub>reference</sub>) and CPD(D<sub>measured</sub>). The corrected D value is the analytical correction based on the size of the data set.

Figure 2.13 (b) – Completion time reliability analysis for task and subtask level of a clip application. Horizontal difference between CPD(D<sub>reference</sub>) and CPD(D<sub>measured</sub>). (TT: total time, AP: approach, TM: tissue manipulation, CP: clip application, WD: withdrawal).
Figure 2.14 (a) – Joint angle distribution reliability analysis for the task level of a clip application. KS (D) value between CPD(D_{reference}) and CPD(D_{measured}). The corrected D value is the analytical correction based on the size of the data set.

Figure 2.14 (b) – Joint angle distribution reliability analysis for the task level of a clip application. Horizontal difference between CPD(D_{reference}) and CPD(D_{measured}).
Figure 2.15 (a) – Similarity sequence of clipping subtasks. KS (D) value between $\text{CPD}(D_{\text{reference}})$ and $\text{CPD}(D_{\text{measured}})$. The corrected $D$ value is the analytical correction based on the size of the data set.

Figure 2.15 (b) – Similarity sequence of clipping subtasks. Horizontal difference between $\text{CPD}(D_{\text{reference}})$ and $\text{CPD}(D_{\text{measured}})$.
2.3.3 Context Comparisons

Contextual comparisons of completion time and kinematic performance measures were evaluated at the task and subtasks levels for the clip application task. Using the K-S statistic we compared odd vs. even; first half vs. second half (1st vs. 2nd) of the clip applications. We also investigated separation of the cystic artery vs. the cystic duct (CA vs. CD); control of bleeding vs. vessel separation (bleed vs. VS); and first placement vs. subsequent placement (init. vs. sub.) on the same vessel. Figure 2.16a illustrates the clipping task and subtask completion time difference scores. For each comparison with the exception of bleed vs. VS we notice the lower bound is zero which suggests we should notice differences in completion time for clips applied for the purpose of vessel separation and clips applied to control bleeding. For the remaining kinematic performance measures we present data only from the task level, as there is little difference between difference measures for each subtask. Figure 2.16b and 2.16c show the difference measures for kinematic performance measures and the mean membership action states, respectively.

![Figure 2.16 (a) - Kolmogorov-Smirnov difference statistic of completion time performance measures for the tasks and subtask levels in different contexts. The confidence intervals marked indicate 5% and 95% confidence and are based on a bootstrap estimate.](image)

*Figure 2.16 (a) - Kolmogorov-Smirnov difference statistic of completion time performance measures for the tasks and subtask levels in different contexts. The confidence intervals marked indicate 5% and 95% confidence and are based on a bootstrap estimate.*
Figure 2.16 (b) – Kolmogorov-Smirnov difference statistic of kinematic performance measures in different contexts. The confidence intervals marked indicate 5% and 95% confidence and are based on a bootstrap estimate.

Figure 2.16 (c) – Kolmogorov-Smirnov difference statistic of mean membership action variables. The confidence intervals marked indicate 5% and 95% confidence and are based on a bootstrap estimate.
2.3.4 Clinical Interpretation

The results of the performance measure reduction are shown in Figure 2.17 for each of the seven procedures with position data. To improve clarity the confidence bounds are not shown. Completion times appear to be influenced by procedural variability more than the postural and sequencing measures.

Figure 2.17 – Performance measure reduction for a clipping task for seven procedures. Performance measures from each procedure are compared to a reference group consisting of all other performance measures. The average performance measure value for each performance measure type (completion time, postural, and sequencing) (left). The average averaged performance measure for each procedure (right).

2.4 Discussion

The primary purpose of this study was to present a method for evaluating surgeon performance and to demonstrate the reliability of our measurement equipment and our selected performance measures. We also showed the utility of such measurements in making clinical performance assessments. We selected the laparoscopic cholecystectomy as our baseline procedure for study, as it is the most common and well-defined minimally invasive procedure (Traverso 1997), and used an optoelectronic measurement system because it is generally well tolerated by surgeons. We also selected the clipping task as a more specific example as clips are frequently applied (in the order of 7 per procedure) during a standard procedure.
2.4.1 Data Acquisition

Our primary contribution was the development of a methodology for quantifying and comparing surgeon performance in different contexts. To achieve this goal, we used an optoelectronic motion capture system and video to collect tool kinematic data, joint angle position, event sequencing, and time data. Unfortunately, we found limitations in our instrumentation, which prevented us from investigating all aspects of performance. We did however collect a considerable amount of useful kinematic, joint angle, sequencing and time data. We found that during tissue manipulation tasks (as opposed to instrument exchanges or maintenance), sampling rates were roughly 20 Hz and joint angle markers were fully visible 80% (SD: 10%) of the time; when these markers did become obscured or otherwise unreadable, the resulting gaps in the data typically lasted for only a fraction of a second (we obtained postural data at least one reading per second 92% (SD: 4%) of the time). This sampling rate is perfectly adequate for postural analyses (Bhattacharya et al., 1999). In combination with postural data, we successfully recorded tool tip position 82% (SD: 12%) (after interpolation) of the time the surgeon was performing clipping tasks. Using a generalized cross validation interpolation technique, we were able to interpolate half-second gaps with estimated RMS errors less than 1.0 mm. Despite our efforts to track continuous tool tip data, we found our data to have sizable missing data segments because of marker occlusions making it difficult to extract useful kinematic performance measures; thereby reducing our ability to establish reliability.

2.4.2 Hierarchical Decomposition

In our comparative analyses, we found the hierarchical decomposition useful as an organizational tool for evaluating and comparing performance measures for the phase, stage, task and subtask levels. Based on work done in the lab, we found it difficult to automatically extract action states based on our kinematic data. The action membership functions often suggest membership in more than one state at any particular instant making it difficult to detect the duration of action states. Our method of extracting kinematic data is an improvement over Cao's video based system and shows limited potential for reliably detecting action states automatically. In future studies, the
hierarchical decomposition will be beneficial for relating various aspects of surgery to simulators since many bench top systems are designed for specific surgical tasks rather than the entire procedure (Derossis 1998).

2.4.3 Reliability Analysis

2.4.3.1 Procedural Specific Variability

We used a modified K-S analysis technique to investigate the effect of patient variability over multiple trials. The purpose of this analysis was to show evidence that procedural specific variability is significant and to estimate the extent of inter-procedural variability. We measured the vertical distance (using the K-S approach) and the horizontal difference in means between the two difference measure ($D$ and $\delta$) distributions. The difference measure distributions represent procedural variability ($D_{\text{measured}}$) and our best estimate without procedural variability ($D_{\text{reference}}$).

Using the K-S statistic we found evidence of procedure variability in all our comparisons of performance measures. We expect to see a certain degree of procedure variability and by sampling more procedures the variability of our performance measures should decrease. We found the confidence intervals of the $D$ estimates decreased and as expected all upper bound limits decreased with increasing number of procedures considered, with the exception of the joint angle distribution comparisons. The mean corrected $D$ values displayed some rather counterintuitive results as in many cases its value did not consistently decrease. This result is difficult to explain and is possibly the result of our approach to correcting $D$ values for sample size. The empirically based correction equation is intended for comparing equally sized distributions and has not yet been verified for unequal sized distributions. In our case we used the effective $N_e = N_1 N_2 / (N_1 + N_2)$ for non-equally sized distributions. Another possible explanation for the non-decreasing behavior of the mean corrected $D$ values is the lack of data. In comparing these distributions using the horizontal difference in means we notice a general decreasing trend with increasing number of procedures considered. In some cases we notice a negative difference ($D_{\text{measured}}$ is to the left of $D_{\text{reference}}$), which we interpret as a random resampling effect.
The large differences in our joint angle difference distributions can be explained by our approach in using distributions of joint angles rather than an individual performance measure for each clipping task. In calculating $CPD(D_{\text{reference}})$ we use a large collection of joint angles across seven procedures. The resampling process produces $CPD(D_{\text{reference}})$, where the confidence bounds and mean value decrease with increasing number of procedures. The mean value of $CPD(D_{\text{reference}})$ is small compared with $CPD(D_{\text{measured}})$ which is to be expected since we compare distributions which are essentially the same. The large values of $D_{\text{ref, meas}}$ are expected, as these distributions are certainly different. The confidence bounds on $D_{\text{ref, meas}}$ are all decreasing and in all cases the difference in means of the $D$ distribution is decreasing. The range $CPD(D_{\text{measured}})$ values are small (~0.2-0.3) which suggests these are appropriate measures of performance.

Overall, we found generally the extent of inter-procedural variability is reduced as shown by the decreasing trend of the upper confidence bounds and differences in distribution means for the increasing number of procedures analyzed suggesting an improvement in our resolution to detect differences in performance. The tradeoff of reducing variability is that we need to examine more procedures which is a time consuming and costly process. To our knowledge there has been no attempt to quantify surgeon performance in the OR (only subjective scoring evaluations) or quantify the effect of procedural specific variability. Performance evaluations in the OR have traditionally been done through subjective evaluation, which have "poor reliability and unknown validity" (Lentz, 2001). Our method shows promise as an effective way to quantitatively assess surgeon performance.

Our reliability analysis of the completion time, joint angle and sequencing performance measures suggests these are reliable measures of performance. From our analysis of event sequencing it is unclear whether the results suggest the surgeon is extremely consistent or the event sequence is inherent to the way a procedure is completed. At particular levels within the hierarchy we suspect event sequencing is dependent on the structure of the procedure rather than a measure of skill. Further measurements of surgeons with differing skill levels is required to make this evaluation. Perhaps a more appropriate model for our
application would allow us to highlight deviations from an ‘ideal’ procedural sequence or incorporate details of past events. We found our sequencing model to be generally well suited for describing high-level differences between procedures in contrast to Rosen’s Markov Modeling approach which is appropriate for characterizing low level state transition rates as illustrated on low-medium-high force torque states (Rosen, 2001). One potential limitation of Rosen’s sequencing work is the inability to reveal high-level details of surgeon performance. Together, Markov Modeling and our event sequencing method can coexist and be useful for identify differences in performance.

2.4.3.2 Context Comparisons

The K-S statistic was used to test for differences in performance measures in different contexts in order to demonstrate its utility. We discovered negligible differences between the time measures in the odd vs. even and 1st half vs. 2nd half clips suggesting there are no obvious learning effects and little effect of patient variability. The completion time difference measures are on the order of 0.2 on the difference scale, which suggests time is a reliable measure of performance. Our kinematic measures at this point appear to be unreliable as the upper bound of the context comparisons are in the order of 0.9 on the difference scale, which implies to detect differences we must see a difference greater than 0.9. Therefore our current kinematic measures show poor reliability. More data is required to identify if kinematic measures are reliable for distinguishing differences in performance. In future studies it might be useful to consolidate data across comparable tasks in different contexts (ie. find performance measures for all tool approaches, not just clipping). There appears to be a slight difference between clips applied to control bleeding and clips applied for vessel separation. This result provides insight into how we can develop new simulators as we may wish to train surgeons differently for clips applied in the context of vessel separation than for controlling bleeding. Overall the time and ergonomic measures seem to show most promise for distinguishing differences.
2.4.4 Clinical Utility

We presented an approach to consolidate performance measures which is important for assessing global performance. Consolidating completion time, postural and event sequencing performance measures from each procedure, we found that completion time is most affected by procedural variability. Our postural and event sequencing appear to be most reliable. Further work is required to establish appropriate weighting measures for the reliability of these and other performance measures. Appropriate weighting measures will be based on reliability and importance of each measure in assessing performance.

To establish the resolution of our measurements in distinguishing differences among surgeons, consider comparing the performance of a new surgeon to our current database of a single expert surgeon. If we evaluate the new surgeon on any arbitrary patient in the OR we would establish a new distribution for any particular performance measure \((CPD_{\text{spec}})\), for example time. Comparing the \(CPD_{\text{spec}}\) and \(CPD_T\) (from our expert surgeon) we get a difference measure \(D_{\text{spec}}\); if \(D_{\text{spec}}\) is less than the 95\(^{\text{th}}\) percentile of \(CPD(D_{\text{measured}})\), then we conclude the performance of the new surgeon is essentially the same (See Figure 2.18). We use the 95\(^{\text{th}}\) percentile of \(CPD(D_{\text{measured}})\) as our critical value \((D_{\text{crit}})\) for distinguishing differences among surgeons.
We can also convert back from the non-dimensional $D$ values to dimensional values, which may be more familiar to a surgeon or surgical educator. Consider using the $D_{cr}$ value for establishing how much of a particular performance measure value is required to detect a difference. To make this quantification, we simply shift the $CPD_T$ until the $D$ value between the shifted and nonshifted $CPD_T$ is equal to $D_{cr}$. The amount of shift is correlated to the difference in performance measure value. Figure 2.19 shows the resolution of our time measurements for detecting differences across multiple procedures. From this demonstration we notice our resolution improves over multiple procedures, although the rate of improvement decreases with increasing procedures. It appears the analysis of two procedures is most appropriate at this point for establishing measures of difference.
Figure 2.19 – Resolution of completion time in detecting difference for increasing number of procedures analyzed.

2.4.5 Future Work

The long-term goal of this work is to develop a validated simulation for applications in training, performance evaluation and equipment design. The value of this simulation depends on how closely it captures the elements of perceptual-motor skills used in an actual task (Sanders 1991). We plan both to develop simple desktop simulations (perhaps based on existing ones – e.g., Derossis 1998) which demonstrate tool and arm motions similar to those used in actual surgery by our expert surgeon and to obtain access to commercial simulator units. Using our data collection techniques (where we expect to see acquisition improvements in the controlled setting of a simulator) and statistical analysis procedures, we will establish the reliability of our performance measures in a simulated setting. We also need to measure performance of surgeons at various skill levels in both the OR and simulator to develop a database of performance measures at various skill levels. Using this collection of data we would hope to show that on average our inter-procedural variability is small compared with inter-surgeon performance differences. In selecting performance measures, we are considering investigating the distribution of
performance measures rather than individual performance measures for different levels of the decomposition (ie. distribution of velocities vs. average velocity) because we want to increase reliability of kinematic measures.

2.5 Conclusion

We used an automated high-frequency tool tracking and postural measurement system to identify kinematic features of specific surgical actions of laparoscopic cholecystectomies as defined in a hierarchical decomposition diagram. The results demonstrate that an optoelectronic and video motion analysis, particularly in conjunction with a missing data interpolation technique, is appropriate for postural measurements and reasonable for recording kinematic data in the operating room, despite occasional periods of missing data.

Based on the feasibility of using an optoelectronic motion capture system we developed a methodology for validating surgical simulators and classifying surgical performance. The methodology is based on making comparisons of performance measure distributions in difference contexts using the K-S statistic. For the most part, our comparisons of performance measure distributions showed consistent reliability over time by a trained surgeon, and so are likely reliable measures of performance. The kinematic measures require further comparative analysis and more data to show reliability.

From our inter-procedural reliability analysis we showed the significance of procedural variability and found as the number of evaluated procedures increases the resolution of our ability to detect differences improves. The presented reliable measures and data will form the basis for quantifying surgical skill, and should be useful in validating surgical simulations for use in training, certifying surgeons and designing and evaluating new surgical tools.
Chapter 3
Repeatability of Biomechanical Stresses in Laparoscopic Surgery

3.1 Introduction

The use of minimally invasive techniques has become widespread over the last decade. Improvements in surgical instrumentation have made it possible for surgeons to undertake increasingly complex laparoscopic surgical procedures. Unfortunately the effects of these instruments on surgeon performance, comfort and safety have not been well evaluated. Minimally invasive surgeons are often forced to use uncomfortable and potentially harmful postures, which may lead to long-term strain injuries and musculoskeletal disorders such as Laparoscopic Surgeon’s Thumb (Horgan 1997) and carpal tunnel syndrome (Buschbacher 1994). To identify and rectify such problems, ergonomists will typically assess working postures using either direct observation or instruments such as goniometers or motion analysis systems (Genaidy 1994). Some researchers have looked at the joint forces of sustained and repetitive postures (Viikari-Juntura 1999). Deviations from neutral posture have been correlated with injury risks, so ergonomic stress ratings can be assigned to the history of joint angles observed during a task in a surgical setting. The Rapid Upper Limb Assessment (RULA) system, which was designed for light manual tasks, is an appropriate scoring system for evaluating biomechanical stress of a surgeon (McAtamney 1993).

Ergonomic assessments in minimally invasive surgery are rare, despite widespread acknowledgement that strain injuries do occur to surgeons. Postural assessments are common in repetitive task workplace settings, but only a small number of studies have been done in the operating room (OR). Radermacher et al. (1996) was one of the first to assess minimally invasive surgeries and outlined improvements for workplace design. Berguer showed the postures in laparoscopic surgery are more stressful than in conventional open surgery (Berguer 1997) and that instrument design affects fatigue
levels (Berguer 1999). In these past studies, visual or video based observations have formed the basis of postural assessment. The effect of patient and procedural variability on a surgeon stress remains poorly understood.

In this study we expand the efforts of Person’s “proof of concept” pilot study, which demonstrated the feasibility of using an optoelectronic and video motion analysis system to perform a continuous ergonomic posture analysis in the OR (Person, 2000). In this research we investigate the performance and postural stress of an expert surgeon over multiple procedures. The objective of this study is to provide an assessment of biomechanical stresses experienced by an expert surgeon over seven laparoscopic cholecystectomy procedures. More specifically we investigate the effect of patient and procedural specific factors on measurement of biomechanical stress and discuss if a single procedure provides a reliable estimate of stress experienced by a surgeon. This objective is met by investigating similarity of joint angle distributions and by identifying similarity of biomechanical stress scores.

3.2 Methods

3.2.1 Equipment

Refer to Section 2.2.2 in Chapter 2 for information on experimental protocol and the data acquisition system.

3.2.2 Kinematic Arm Model, Calibration and Joint Angle Calculations

The Polaris system reports the spatial position of the marker arrays. Joint angles are calculated from the measured position of the marker arrays located on the surgeon’s hand, forearm, upper arm and torso. We used a simplified seven degree-of-freedom (DOF) rigid-body model of the upper limb. The model allows us to represent shoulder abduction/adduction, flexion/extension, and internal/external rotation, elbow flexion/extension, pronation/supination and wrist flexion/extension and radial/ulnar deviation. Using this model, we applied a kinematic calibration method based on circle
and sphere fitting techniques to locate the approximate centre of each of the shoulder, elbow, and wrist joints. The method locates a joint axis or centre of rotation by fitting a circle (or sphere) to a cloud of raw data points collected from limb markers and expressed in a reference frame attached to an adjacent limb segment. These techniques are well known in the biomechanics and robotics literature (e.g., Halvorsen 1999). We used the joint angle definitions provided by Anglin (1993) to ensure consistency with the angles used in the RULA scoring system.

3.2.3 Comparative Analysis (KS Statistic)

We would expect to see some degree of difference in joint angle distributions as a result of patient, procedural, and measurement variability, however since the surgeon is fully trained we expect that his approach to each patient will be as consistent as possible. We would hope that there is sufficiently little variation between procedures that measurements obtained during a single procedure will be reliable indicators of the physical stress experienced by the surgeon. To demonstrate the repeatability of stress scores for a surgeon over multiple surgeries, we report the differences between the distributions of joint angles obtained in each procedure taken singly and those obtained from the remaining six surgeries. The Kolmogorov-Smirnov (K-S) statistic, $D$, provides a measure of difference between two distributions and indicates the maximum vertical difference between pairs of cumulative distribution functions as shown in Figure 3.1; $D$ ranges from 0 (similar) to 1 (different). Using a bootstrapping approach outlined in Appendix E we establish a 90% confidence bound on our estimate of $D$. The associated $p$ value expresses the probability that the two measured distributions arise from the same underlying distribution.
Figure 3.1 - The Kolmogorov-Smirnov statistic, $D$, is defined to be the maximum vertical difference between two cumulative probability distribution. The cumulative distribution function for a finite data set is represented as a staircase.

The difference measure calculated from the K-S statistic for each joint of each procedure provides a way of quantitatively assessing the difference between two distributions. Since we performed this experiment using the same tool set over seven procedures the difference measure essentially provides us with a benchmark for that particular tool set. In future analyses we may be interested in evaluating new tool designs and their effect on surgeon posture. A comparison of the difference measures from the new tool design and benchmarked values of the old tool set will provide insight as to how well the new tool influences surgeon posture. For example, consider developing a tool to reduce stress in the wrist. Our goal would be to design a tool, which minimizes the frequency of wrist flexion/extension and ulnar/radial deviation. If we collect joint angle data of a surgeon using this new tool and compare these distributions with our database of joint angle measures (old tool design and the same surgeon) we would expect to see a difference in distributions. The K-S statistic provides a means of quantitatively assessing the difference between the two distributions of joint angles and from this we quantify the improvement of the new tool.

3.2.4 Ergonomic Stress Scoring System

To rate the stress experienced by a surgeon we used a modified version of the Rapid Upper Limb Assessment (RULA) technique (McAtamney 1993). The Ovako Working Posture Analysis System (OWAS) for postural evaluation was considered but deemed inappropriate for our application (Karhu 1997). The OWAS technique was originally
designed for assessment of workers in heavy industry and would require significant modifications to be appropriate for our applications. The RULA technique is a posture evaluation method designed for assessing upper limb postures during light manipulation tasks. Stress exposure is evaluated using body posture diagrams and scoring tables, which specify different posture zones shown in Figure 3.2. The posture limit guidelines and scores are validated and based on the findings of several ergonomic studies (McAtamney 1993). We used a modified version of the RULA scoring system, which incorporated continuous scoring measures to improve the sensitivity of our assessments. The RULA scoring system is based on a discrete score for selected joint angle regions. Using a continuous scoring system we are able to more accurately define stress for a range of joint angles. Since the standard RULA scales have different ranges for each joint, we normalized all scores such that 0.0 corresponds to the minimum RULA value for a joint and 1.0 corresponds to the maximum RULA value.

![RULA Posture Zones](image)

*Figure 3.2 – R.U.L.A. Posture zones. Zero reference positions are also shown.*
3.2.4.1 Modified Discrete RULA Method

We modified the RULA scoring system slightly to resolve the ambiguity in the scheme's use of the clinical terms 'abduction/adduction' and 'flexion-extension'. We adopted the angle convention used by Keyserling (1986), where shoulder flexion/abduction is classified by identifying elbow elevation <45° from the neutral (anatomical) position as neutral, 45°-90° as mild elevation, and >90° as severe elevation. A score similar to the RULA method is applied to indicate the posture level: neutral (1), mild elevation (2), severe elevation (3).

We also modified the wrist flexion-extension and radial-ulnar deviation penalty zones. In the original RULA system, any non-neutral wrist postures (assessed visually) are considered stressful, but since measurements from a motion analysis system will always contain some noise, we must choose a non-zero finite range to represent neutral posture; we have therefore defined neutral wrist angles as those lying within ±5° of the neutral position. Finally, in order to numerically classify the range of forearm pronation/supination, we adopted the posture classification presented by Genaidy (1994), where pronation or supination outside a ±15° range from neutral is considered poor posture.

3.2.4.2 Modified Continuous RULA Method

The modified discrete RULA method described in the previous section provides discontinuous parameter jumps in the RULA Score separating each posture zone. The original RULA system was designed for visual or video observation where continuous joint angle measurements are not available. The actual stress on the soft tissues and the consequent risk of injury will be a continuous function of the joint angles. A more appropriate scoring model would incorporate the same scoring approach by replacing the discontinuities with continuous transition functions. This in effect represents a more realistic method of scoring stress. For example, if we consider flexion of the wrist, 5.1° of wrist flexion is only slightly more stressful than wrist flexion of 4.9°. Using the discrete
modified RULA method we see completely different stress scores for each of the two joint angles. Using a continuous scoring system we would see a much smoother transition, which is more representative of joint stress. Figure 3.3 shows the discrete scores and continuous transition scoring functions chosen for the modified RULA scoring system of each joint.

![Figure 3.3](image-url)

*Figure 3.3 – Discrete and continuous modified RULA posture scores adopted for current study of continuous ergonomic posture analysis during minimally invasive surgery.*
3.2.5 Post-Operative Calculations

Post-operatively, we calculated the joint centres relative to the attached marker arrays and calculated joint angles and the associated discrete and continuous RULA scores at each sampling interval throughout the procedure. Using the correlated video data we reviewed the entire procedure and divided the procedure into phases based on a hierarchical decomposition outlined in Chapter 3. The three phases of interest include the: cystic duct dissection, gallbladder dissection, and gallbladder removal as classified by Traverso et al. (1997). For each major phase of the surgery, we computed the fraction of time the surgeon spent at each stress level on the discrete RULA scale. We also compared these distributions across phases of the surgery to determine which phases were most stressful.

Using these normalized continuous RULA values, we compute a normalized weighted postural stress score (NWPSS) for a period of observation by integrating the instantaneous scaled RULA scores (sRULA(t)) over the observation period, T, and then dividing by the period length to produce an average score:

\[ NWPSS^j = \frac{1}{T} \int_0^T sRULA^j(t) dt, \]

where the superscript \( j \) refers to a particular joint.

3.3 Results

3.3.1 Marker Visibility

A certain amount of position data is lost during the course of the surgery due to marker occlusion or internal localization errors. (e.g., Polaris is unable to find a marker within 0.5 mm of its expected location). These data were excluded in determining the total percentage spent in a given posture. Figure 3.4 illustrates the marker array visibility averaged over the seven procedures. On average, 36% (SD: 7%) of all joint angle data were missing during the entire surgery over seven trials (before interpolation).
Figure 3.4 - The values labeled on each marker represent the percentage of time it was tracked over all seven procedures after interpolation.

We used a generalized cross validation (GCV) filtering technique to filter, interpolate and resample position data at a constant 20 Hz sampling interval (Woltring 1986). We were able to reduce the amount of missing joint angle data by 20% (SD: 2%) using our interpolation technique with a maximum 1mm RMS error across a maximum gap size of 0.5 sec (10 samples).

Using the correlated video images we separated manipulation from non-manipulation tasks and found that the majority of missing data could be attributed to marker occlusions during segments not directly related to performing a surgical task (e.g. defogging the laparoscope lens, changing instruments, etc.). During tissue manipulation tasks and after interpolation, we had complete joint angle data 80% (SD: 10%) of the time and at least one sample per second 92% (SD: 4%) of the time.

3.3.2 Comparative Analysis

For each joint, we calculated and plotted the cumulative probability distribution (CPD) of joint angles during each procedure as shown in Figure 3.5. The CPD is useful to visualize the distribution of the joint angle data and differences among individual distributions.
Figure 3.5 - Graphical representation of cumulative probability distributions of each joint and for each procedure. Each curve in each plot represents the joint angle distribution of one procedure.

3.3.2.1 Measurement Errors

In our system, measurement errors can be attributed to: (1) internal localization errors, (2) marker shifts, and (3) neutral posture reference. The internal localization errors are inherent to the system and cause positional and angular orientation errors in the order of
0.35 mm and 1.1° respectively. These errors show up as noise in the position and orientation signals and are filtered out using the GCV filter described in Section 3.4.3.2.

Marker shifts occur when a marker is bumped into a different position. Great care was taken to position and align the markers in such a way they are not disrupted during the procedure. Only one documented marker shift occurred in all of the procedures analyzed. In procedure 4 the hand marker shifted at minute 6 (total procedural time 44 min) which was the result of poor marker adhesion to the hand. The profile of the CPD in Figure 3.6 shows evidence of a shift in the wrist RU angle distribution. We are currently developing algorithms based on an automatic joint centre identification method reported recently by O'Brien which will help compensate and correct for marker shifts (O'Brien 2000).

**Figure 3.6** - Graphical representation of the cumulative probability distributions of wrist ulnar / radial deviation. Procedure 4 shows evidence of a marker shift.
The neutral posture reference taken at the end of each procedure is somewhat subjective in that the surgeon simply is asked to adopt a neutral posture. The neutral posture reference is the largest source of error for each joint angle measurement. The data collected for the neutral position is used as the absolute reference for all joint angle calculations for a particular procedure. We conducted an in-lab experiment to investigate the variability of the neutral posture reference position. A subject with no prior knowledge of the system was chosen to repeat a neutral posture reference position 30 times. For each trial the posture was recorded and the joint angles calculated. Since this subject repeated the neutral postures at short intervals, the results are probably better than those of the surgeons, who adopted these postures at intervals of up to several weeks. Figure 3.7 shows the variability of our neutral posture measurements found experimentally in the lab and from the OR data.

![Figure 3.7](image.png)

*Figure 3.7 – Variations in neutral posture measurements found experimentally in the lab and from the OR data.*
3.3.2.2 Distribution Profiles

The CDF of each joint angle has its own unique profile and horizontal shift. The profile (or shape) itself provides insight into the characteristic behavior of each joint. The profile suggests information about frequency of different angular positions during the procedure. Due to variability in defining the neutral posture for each joint, we shifted each distribution such that they share a common mean. The differences in the distributions are therefore primarily due to differences in shape rather than mean values. Figure 3.8 shows a comparison between the joint angle distributions of the elbow before and after the shift and Figure 3.9 shows a summary of all shifted joint angle distributions. The KS (d) statistic is applied to the collection of shifted joint angle distributions.

![Graphical representation of the cumulative probability distributions of elbow flexion and extension (L) Non-shifted data (R) Shifted data (Normalized about the mean of all distributions).](image)

**Figure 3.8** - Graphical representation of the cumulative probability distributions of elbow flexion and extension (L) Non-shifted data (R) Shifted data (Normalized about the mean of all distributions).
Figure 3.9 - Graphical representation of normalized cumulative probability distributions of each joint and for each procedure. Each distribution is normalized to the collective mean of joint data from all procedures. The bottom right figure shows the joint angle distribution shifts for each joint of each procedure.
3.3.2.3 Patient / Procedure Factors

Each procedure introduces a certain amount of variability into the data set. The same expert surgeon was evaluated in all procedures and each case involved a new patient and operating staff. Table 3.1 summarizes the patient and procedural factors which influenced the outcome of each procedure and Figure 3.10(a) shows distinct profile differences for the joint angle distributions of the wrist. Comparing the distribution profiles to the patient and procedural factors outlined in Table 3.1 shows little correspondence between distribution profile and procedural variability.

Table 3.1 - General comments regarding each procedure.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Patient Factors</th>
<th>Procedural Factors / Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>2</td>
<td>Heavy</td>
<td>Moderate</td>
</tr>
<tr>
<td>3</td>
<td>Heavy</td>
<td>Normal</td>
</tr>
<tr>
<td>4</td>
<td>Heavy</td>
<td>Difficult</td>
</tr>
<tr>
<td>5</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>6</td>
<td>Heavy</td>
<td>Normal</td>
</tr>
<tr>
<td>7</td>
<td>Heavy</td>
<td>Normal</td>
</tr>
</tbody>
</table>


Figure 3.10(a) - Graphical representation of cumulative probability distributions correlating procedures to groupings of joint angle distributions. Normalized cumulative probability distributions of the wrist joint for each procedure (left - wrist flexion/extension, right - wrist radial/ulnar).
3.3.2.4 Difference Measures

The difference measure $D$ calculated from the K-S statistic for each joint of each procedure provides a way of quantitatively assessing the difference between two distributions. Plotting the $D$ values for each procedure on a difference scale from 0 to 1 gives us an idea as to the variability of these distributions for an expert surgeon using a standard toolset in a laparoscopic cholecystectomy. The position of the $D$ values on the difference scale provides context to the scale itself. In other words we have a known benchmarked difference scale from which further comparisons can be made. Figure 3.11 shows the range of $D$ values for each joint resulting from comparing joint angle

Figure 3.10(b) - Graphical representation of cumulative probability distributions showing types of variability from procedure to procedure. Horizontal shifts of the distributions (left) and differences in the shape of the distributions (right).

In Figure 3.10(b) we notice two types of variability; (1) horizontal shifts of the distributions and (2) differences in the shape of the distributions. Both the horizontal shifts and differences in shapes of distributions suggest patient or procedural variability. The horizontal shifts may also indicate measurement errors. In particular, these measurement errors are likely caused by variability in the neutral posture reference.
distributions from each procedure to the concatenated distribution of the remaining six procedures, along with the average $D$ value and standard deviation for each joint for shifted and non-shifted data sets. The most proximal joints exhibit the least variability ($D \sim 0.10-0.13$, on average), while the wrist joint exhibits the most ($D \sim 0.2$, on average). Since the surgeon is fully trained, these values represent the minimum variations we might expect to find from procedure to procedure when using the same surgical tools.

**KS (D) Statistic for Shifted and Non-Shifted Data**

![KS (D) Statistic for Shifted and Non-Shifted Data](chart)

**Figure 3.11** – Benchmarking the difference measures using the K-S statistic on a difference scale.

### 3.3.3 Posture Scores

Joint postures are calculated for the three main minimally invasive phases of the standardized laparoscopic cholecystectomy operation as defined by Traverso et al. (1997): cystic duct dissection (CDD), gallbladder dissection (GBD), and gallbladder removal (GBR). Table 3.2 lists the time spent in each of these phases by our surgeon, as well as typical times required.
Table 3.2 – The average time taken for the major surgical tasks performed in seven laparoscopic cholecystectomies and compared with typical times from Traverso (1997).

<table>
<thead>
<tr>
<th>Operative Task</th>
<th>Time (mins)</th>
<th>Typical Time (mins ± std. dev.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total operation time</td>
<td>68.7± 11.0(100%)</td>
<td>72 ± 28 (100%)</td>
</tr>
<tr>
<td>Cystic duct dissection</td>
<td>18.8±5.5 (27.5±7.1%)</td>
<td>15 ± 11 (21%)</td>
</tr>
<tr>
<td>Gallbladder dissection</td>
<td>15.5±4.8 (22.3±4.8%)</td>
<td>14 ± 8 (19%)</td>
</tr>
<tr>
<td>Remove gallbladder</td>
<td>8.7±1.8 (12.8±2.9%)</td>
<td>12 ± 6 (17%)</td>
</tr>
</tbody>
</table>

We computed the modified discrete and continuous RULA posture scores illustrated in Figure 3.3 for the three operative phases (CDD, GBD, and GBR). Figures 3.12, 3.13 and 3.14 show the average contribution of the discrete ergonomic stress levels for each shifted joint angle distribution of interest during the three major phases of the surgery. A normalized weighted postural stress score (NWPSS) is calculated for each phase of the procedure for both the discrete and continuous scoring systems. For each joint there is lower variance using the continuous scoring system versus the discrete system. The NWPSS provides a directly comparable measure of overall postural stress in each joint.

Shifted Data - Stress contribution level: Cystic duct dissection

<table>
<thead>
<tr>
<th>Continuous NWPSS</th>
<th>13.8±2.5%</th>
<th>56.5±7.3%</th>
<th>54.9±12.3%</th>
<th>71.5±10.1%</th>
<th>32.1±16.4%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete NWPSS</td>
<td>7.8±5.2%</td>
<td>68.6±8.8%</td>
<td>50.9±11.9%</td>
<td>69.4±11.8%</td>
<td>27.9±19.1%</td>
</tr>
</tbody>
</table>

Figure 3.12 – Cystic duct dissection (CDD) stress levels for each joint angle normalized and averaged over seven procedures, including the continuous and discrete NWPSS.
Shifted Data - Stress contribution level: Gallbladder dissection

<table>
<thead>
<tr>
<th></th>
<th>Continuous NWPSS</th>
<th>Discrete NWPSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12.2±0.7%</td>
<td>4.6±1.8%</td>
</tr>
<tr>
<td></td>
<td>39.5±8.5%</td>
<td>39.7±11.9%</td>
</tr>
<tr>
<td></td>
<td>60.2±16.4%</td>
<td>57.6±18.1%</td>
</tr>
<tr>
<td></td>
<td>65.7±14.2%</td>
<td>62.6±17.9%</td>
</tr>
<tr>
<td></td>
<td>59.9±20.5%</td>
<td>57.1±25.4%</td>
</tr>
</tbody>
</table>

Figure 3.13 – Gallbladder dissection (GBD) stress levels for each joint angle normalized and averaged over seven procedures, including the continuous and discrete NWPSS.

Shifted Data - Stress contribution level: Gallbladder Removal

<table>
<thead>
<tr>
<th></th>
<th>Continuous NWPSS</th>
<th>Discrete NWPSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12.3±3.1%</td>
<td>6.9±4.5%</td>
</tr>
<tr>
<td></td>
<td>43.3±5.3%</td>
<td>50.5±7.3%</td>
</tr>
<tr>
<td></td>
<td>56.8±18.2%</td>
<td>53.8±18.7%</td>
</tr>
<tr>
<td></td>
<td>73.8±7.4%</td>
<td>72.2±8.5%</td>
</tr>
<tr>
<td></td>
<td>32.8±10.8%</td>
<td>28.9±12.3%</td>
</tr>
</tbody>
</table>

Figure 3.14 – Gallbladder removal (GBR) stress levels for each joint angle normalized and averaged over seven procedures, including the continuous and discrete NWPSS.
3.4 Discussion

The primary purpose of this study was to provide an assessment of the repeatability and level of biomechanical stresses experienced by an expert surgeon over multiple laparoscopic cholecystectomy procedures and to establish the foundation of a method to quantitatively assess new tool designs. We selected the laparoscopic cholecystectomy procedure to study as it is the most common and well-defined minimally invasive procedure (Traverso 1997), and we used an optoelectronic measurement system because it is well tolerated by surgeons and reasonably easy to implement.

3.4.1 Tool Tracking

During tissue manipulation tasks, sampling rates were 20 Hz and the markers were fully visible almost 80% of the time; when markers did become obscured or otherwise unreadable, the resulting gaps in the data typically lasted for only a fraction of a second. After interpolation we obtained at least one reading per second 92% of the time. This sampling rate is adequate for postural analyses; Bhattacharya et al. (1999), for example, sampled carpenters' postures at 1 Hz and Person (2001) sampled at 1 Hz, 96% of the time.

3.4.2 Joint Angle Comparison

A comparison of joint angle distributions was used to investigate the effect of procedure specific factors on measurement of biomechanical stress and to test if a single procedure provides a reliable estimate of stresses experienced by a surgeon. An estimate of the similarity of the joint angle distributions is made using the K-S ($D$) statistic. The horizontal shifts of the cumulative probability distributions (CPD) indicate variability caused both by measurement error and patient and procedural influences. The most significant cause of measurement error was the neutral posture reference position. In an attempt to investigate the effect of patient and procedural variability each CPD was normalized to a common mean, providing a means to investigate the shape similarity of
each distribution. The normalized or shifted CPD's allows investigation of the distribution profiles, which most likely illustrates the effect of patient and procedural variability. This may underestimate $D$ but does put a lower bound on expected variability.

The low value of $D$ for the shoulder, elbow, and forearm suggests these distributions are only slightly affected by patient variability. Variance of $D$ for the wrist flexion/extension and wrist radial/ulnar deviation suggest moderate effects of patient and procedural variability. In particular, difficult patient types cause greater differences in $D$ values.

Using the K-S ($D$) statistic to calculate measures of difference among joint angle distributions, we established a point on the scale for future assessments of new tool designs. A quantitative assessment of new tool designs can be evaluated on the difference scale. For example the evaluation of a new tool designed to improve wrist ergonomics would most likely show similar $D$ values for the shoulder, elbow, and forearm and significantly different $D$ values for the wrist joint angle distributions, coupled with a lower NWPSS.

Patient variability had a moderate effect on the joint angle distributions. The joint most significantly affected by patient variability is the wrist in both flexion/extension and radial/ulnar deviation. For future analysis of new tool design it may be appropriate to evaluate performance on a single patient provided the patient is as ‘normal’ as possible and if we require differences in $D$ greater than ~0.2.

### 3.4.3 Biomechanical Stress Levels

The normalized stress (NWPSS) for all joints is used to assess the biomechanical stresses experienced by an expert surgeon over seven laparoscopic cholecystectomies. The averaged NWPSS for all joints is comparatively high throughout all the procedures. Figure 3.15 shows the averaged shifted continuous and discrete NWPSS values. The variance of continuous scores are consistently less than the discrete NWPSS scores which suggests there is a difference in our ability to predict ergonomic stress and the continuous
RULA scoring system is a more appropriate method for evaluating stress. The results also show that the wrist and elbow are consistently stressed in all phases of the procedures, which agrees with other reports (Berguer 1998). The forearm is stressed to a lesser degree but it is noted that the GBD phase has more ergonomic stress than the CCD and GBR phases. During the GBD phase, the spatula tool is used most often and since the spatula has no handle the surgeon is forced to hold the tool in awkward ways resulting in high ergonomic stress during the GBD phase.

Figure 3.15 – Shifted continuous and discrete NWPSS averaged over seven procedures for each joint in each surgical phase.

3.4.4 Future Work

Our motion capture system has proven to be a useful ergonomic analysis tool; however, there are several improvements which could be made in the future. First, develop a constraining jig for the upper limb and arm for the neutral posture reference. This issue may also be addressed by implementing O'Brien's automated joint centre and joint axis locating techniques (O'Brien 2000).

Secondly, we did experience frequent short-term occlusion of the markers, which would prevent us from performing dynamics calculations. This problem could be alleviated
somewhat by positioning the camera directly overhead, using a second camera, increasing the number of faces on each marker array to enhance visibility, or complementing the optical system with magnetic trackers.

Finally, we would like to investigate the effects of fatigue resulting from prolonged near-stationary posture and rapid posture transitions (i.e. fatigue vs. repetition). Since we already have joint position as a function of time this analysis would not required additional OR data.

3.5 Conclusions

We successfully used an automated high-frequency postural measurement system to assess the biomechanical stresses experienced by an expert surgeon over seven laparoscopic cholecystectomy procedures. We notice patient variability primarily affects certain joint angle distributions, in particular the wrist and elbow joints. Ergonomic stress levels are high in all phases of surgery; in particular, the elbow and wrist are significantly more stressed than the shoulder and forearm. Using the K-S statistic we created a difference scale to benchmark the effects of procedural variability. This new benchmarking scale can be used to evaluate new tool designs, validate surgical simulations and assess the performance of surgeons. This quantitative assessment method allows use to identify segments of the surgery which are most stressful to surgeon.
Chapter 4
Conclusions and Recommendations

4.1 Introduction

The goal of the research presented in this thesis was to establish a methodology for quantifying the validity of surgical simulators for the purpose of assessing skill and performance of surgeons and evaluating new tool designs. A technique using a motion analysis system to measure surgeon performance during live surgeries was developed. Seven live procedures were recorded (along with video records of an additional eight procedures) and from this data a range of performance parameters were derived. A method was also developed for showing the reliability of our measurements and our established performance parameters. This system is intended to be used in future studies for the purpose of validating surgical simulations and for developing a database of performance measures of surgeons at various skill levels for surgeon certification.

4.2 Review of Present Research

4.2.1 Motion Capture System Feasibility

We used an automated high-frequency tool tracking and postural measurement system to measure the movements of surgical tools and the posture of a surgeon during seven laparoscopic cholecystectomy procedures. To our knowledge this is the first time such a system has been implemented in the OR for the purpose of performance evaluation. In fact, this is the first time anyone has taken a quantitative approach to evaluating performance in the OR. The results demonstrate that an optoelectronic and video motion analysis, particularly in conjunction with a missing data interpolation scheme, is a reasonable method for recording kinematic data in the operating room, despite occasional periods of missing data. This system is particularly well suited for measuring joint angles at low frequencies. When tracking the tool tip, however, this system often misses sizable data segments, thus reducing our ability to extract useful performance measures. For our analysis purposes, tool tip tracking requires continuous high frequency data acquisition.
Using this motion capture system in the OR environment presented many challenges. Not only did we have to deal with equipment logistics, marker sterilization concerns, and patient scheduling; but it turned out that during the time of our experiments we were faced with a province wide nursing strike. These challenges prevented us from taking measurements on a weekly basis (a typical schedule of our surgeon); instead we collected data, on average, every two weeks for a four-month period. An estimated 25-30 hours of preparation work went into every procedure (equipment checks, marker sterilization, and equipment transport) and at least two researchers were required in the OR to operate the data collection system. In order for the recordings to be successful we required the cooperation of operating staff, the surgeon and essentially flawless operation of our equipment. In general, we found the operating staff to be very helpful and on all days the surgeon was willing to participate in the experiment. Our first experiment was a real learning experience. We found our postural markers worked well for calculating joint angles but our tool trackers needed improvement (Our first experiment used planar marker arrays for tracking the tool handle). The results from our first experiment suggested we needed an improved marker array design which lead to the development of the MDMArray. On one occasion our data acquisition software was improperly shut down resulting in formatting problems of the raw data. A lengthy post-operative data analysis was required to rectify the raw data. Improvements to the software and GUI interface prevented future mishaps. At its present state the feasibility of using this system to evaluate a large number of surgeons in the OR is poor and impractical for commercial applications.

The effects of this system on surgeon performance are difficult to quantify. Our surgeon has expressed concerns over the size and weight of the instrumentation used, suggesting their presence may have affected his overall performance. In all of our recordings the surgeon was advised that if he felt uncomfortable wearing the markers at any time during the surgery, he was free to stop the experiment and remove the markers. In none of our recordings was the experiment prematurely interrupted. However, at the end of three procedures the surgeon commented that the system had a distracting effect and added considerable stress to the procedure. On a number of occasions, the surgeon also
expressed that he suffered from "stage fright". At this point it is difficult to tell whether the instrumentation, the patient, or some other external factor lead to the additional stress experienced by the surgeon. We would assume these distractions would reduce the reliability of our overall performance assessment and would be quantified in our analysis of procedural variability. Additional work is required to minimize the impact of such a system on the performance of a surgeon. Perhaps smaller instrumentation would be of most benefit but ideally we would like to take our measurements without the surgeon even knowing.

At its present state, this system requires a considerable amount of manual and human input. For example, to identify segments of the procedure during which there are manipulation tasks required the researcher to painstakingly review each frame of the video data. This human input not only introduces a certain amount of subjectiveness (although we tried to be as consistent as possible) but also reduces its feasibility to be used in a widespread research study. A considerable amount of time was also spent calibrating the marker arrays and surgical instrumentation. Our surgeon has also expressed considerable opposition toward the postoperative joint calibration procedure (postoperative joint calibration can be replaced with an automated system for joint centre identification). Further work is required to automate this system, thereby allowing us to expand our data collection capabilities.

Finally, there was some concern about the precision of our tool tip location. The tool tip location was found based on the instantaneous position of the tool marker, which is located at the tool handle. A tool tip transformation from the tool marker was found using sphere fitting techniques at the end of each procedure. The precision of the tool tip location is on the order of ±1.5-2.0 mm, depending on the noise of the marker array. The kinematic details of the tool are affected by the positional accuracy of the tool tip which raises the issue of the accuracy of our kinematic measures. The poor reliability of our kinematic measures may be the result of both poor tool tip accuracy and lack of data.
4.2.2 Performance Measure Reliability for the Assessment of Surgical Performance in the OR.

Having established the feasibility of using an optoelectronic motion capture system, we developed a methodology for validating surgical simulators and classifying surgical performance. A hierarchical decomposition was developed to provide an organizational framework for making comparisons of performance measures in different contexts. Based on work done in the lab, we found it difficult to automatically extract action states based on our kinematic data. Our approach used fuzzy memberships to define the membership of each state at any particular instant. We discovered that the membership functions often suggest membership in more than one state at any particular instant making it difficult to detect the duration of action states. Our method of extracting kinematic data is an improvement over Cao’s video based system and shows limited potential for reliably detecting action states automatically (Cao 1996).

Our methodology is based on making comparisons of performance measure distributions in difference contexts using the Kolmogorov-Smirnov (K-S) statistic. The K-S statistic is most commonly used to test the null hypothesis that two sets of data were drawn from the same underlying distribution. In our case, however, we know that the two data sets are obtained from two separate tasks, so we expect the two distributions to be different. Instead, we use the K-S statistic as a measure of the difference between the two distributions and will interpret significant differences as indicators of performance or skill level differences. The K-S statistic is an improvement over traditionally used statistical tools such as the t-test and f-test which test for significance in means and variance respectively (Von Mises, 1964). The K-S statistic was useful because it is nondimensional, which makes it directly applicable to any measured variable regardless of the units used in the measurement, and because it makes no a priori assumptions about the nature of the statistical distributions of the data. The K-S (D) value gives us an idea of how different two distributions are regardless of how they differ, which is important in our application of performance assessment.
For the most part, our comparisons of performance measure distributions showed that the trained surgeon performed consistently over the seven procedures, so the performance measures we investigated are reasonably reliable measures of performance. Because of difficulties in obtaining sufficient tool tip data, we cannot yet conclude that the kinematic measures we measured were reliable.

From our inter-procedural reliability analysis, we showed the significance of procedural variability and found that, as the number of procedures evaluated increases, the resolution of our ability to detect differences improves. The significance of procedural variability was shown for all of our performance measures (time, kinematics, joint angles, and event sequencing). Improving the resolution of our measurements is important for detecting differences between different skilled surgeons. However, the resolution for detecting a certain level of differences is a trade-off to the cost of measuring additional procedures. If we find we need a certain level of resolution and it requires measuring 10 procedures then the system is impractical, as the cost of measuring the 10 procedures is high. Not only is the cost prohibitive but also the subject will probably have considerable performance gains over those 10 procedures, which are not being detected, thus defeating the purpose of our analysis. In our case, for completion time we need about 6 measurements (ie. one procedure) to distinguish a moderate D value of \( \sim 0.3 \). At this point this seems appropriate, although further work is required to see if we can measure differences between surgeons.

### 4.2.3 Repeatability of Ergonomic Measures in MIS Surgery

In this study, we used the same optoelectronic motion capture system to assess the biomechanical stresses experienced by an expert surgeon. This system was the first-ever high frequency automated posture measurement system to be implemented in the OR and stands as a substantial improvement over traditionally based check-list approach. We used a modified version of the RULA technique for evaluating surgeon joint stress. The original RULA scoring system, based on discrete scores, was developed as a visual checklist assessment for postural stress. Since we have high frequency joint angle data we
modified the RULA scores into a continuous scoring system. We found the standard deviation of stress scores across the seven procedures to be lower using the continuous system vs. the discrete system, thus suggesting the continuous system is more appropriate for high frequency joint angle data.

From our analysis we noticed that patient variability affects certain joint angle distributions more than others, particularly the wrist and elbow joints. More data is required to say definitively if specific patient variability affects certain joint angle distributions in particular ways. At this point it is difficult to tell if a surgeon behaves differently (in terms of postures) because of the patient-specific factors (ie. weight, sex, etc...). From our data we found that postural stress levels are high in all phases of surgery, and the elbow and wrist are significantly more stressed than the shoulder and forearm. This quantitative assessment method allows us to identify segments of the surgery which are most stressful for the surgeon. Using this knowledge to identify stressful states in the OR, we can improve instrumentation for surgeons during laparoscopic surgery and recommend changes in posture to improve comfort.

It is important to note that the RULA scoring system neglects to take into consideration the effect of joint fatigue or loading in the joints. The stress experienced by a surgeon holding a particular posture of a prolonged period of time is different from the same surgeon moving in and out of that posture from a neutral position over a longer period of time. Each scenario would give a similar RULA score but the effects of fatigue and joint stress would be different in each case. Future work may involve consulting with ergonomists to identify the effects of joint fatigue and repetitive joint strain. We may also consider correlating the stress experience by the surgeon with forces and tool tip position.

4.2.4 Overall Performance Assessment

Surgeon performance or skill assessments have become an important topic in the minimally invasive surgery research. Research groups involved in developing training programs or simulations always propose some measure of performance. Completion time
is most commonly used as measure of evaluating performance (Chung 1998, Sackier 1998) as it is easy to measure and is directly affected by the task being performed (Sanders 1991). Although simple to measure and compare, task completion time may not sufficiently describe the skill required in performing a task. Research groups have looked at a number of different performance measures including: precision and speed (Derossis 1998), knot quality (Hanna 1997), economy of arm motion (Eman 1999), errors and task times (Hanna 1998), error propagation (Joice 1998), and force-torque signatures (Rosen, 2001). In our research we considered completion time, posture, event sequencing and tool tip kinematics. Any one of these measures may show performance differences, but the question is which measure, or combination of measures, captures the relevant aspects of real surgery. To address this question, we need to measure the performance of surgeons in the OR and in simulators. Using this collection of performance measures, we need to identify which performance measures are reliable at detecting differences between surgeons of different skill levels. Our research suggests for an expert surgeon in the OR our measures of completion time, event sequencing, and joint angle measures are reliable measures for evaluating performance and will be important for the purpose of validating surgical simulations.

At this point we don’t know exactly how good our performance measures will be at differentiating between different groups of surgeons. We may find that two equally trained surgeons have completely different tool tip kinematics but similar end results. In this case, we would say that the kinematic measures are unreliable. However, if the end result is the same then there must be some commonality in the sequence of events that occurred. This would suggest that perhaps the sequencing performance measure would be appropriate for measuring differences in performance. We are now starting to make distinctions between motor and cognitive skill levels. Future work will require having a broad disciplinary research group of engineers, physiologists, surgeons, and ergonomists.
4.3 Future Research Recommendations

4.3.1 Motion Capture Measurements

In consultation with surgeons and other engineers, several recommendations were derived from our motion analysis study in the OR to enhance the reliability of our measurements. The suggestions for future work in data collection are:

- Improve surgeon instrument motion tracking methods by improving on the marker array designs and camera positioning. This is done by developing an optimal geometric marker array design and by identifying an optimal camera position.
- Investigate the feasibility of using multiple motion tracking cameras and the implementation of alternative tool motion tracking instrumentation such as accelerometer/gyro systems (Integrated Micro Instruments) and Shape Tape (Measurand Inc). The advantage of these techniques are continuous tracking of the tool tip and is limited by having additional wires and attachments to the surgical tool.
- Investigate using NCL’s 6-axis force transducer (ATI Industrial Automation) for calculating force torque signatures of surgical tools (Rosen, 2001).
- Develop an automated system for extracting performance measures from video and/or kinematic data.
- Develop a constraining jig for the upper limb neutral posture reference. This will help eliminate some of the distribution shifts of the marker data from procedure to procedure.
- Implement an automated technique for determining the joint centres and joint axes (O’Brien 2000).
4.3.2 Performance Measures and Statistical Analysis

Recommendations for additional performance measures and alternative statistical analysis techniques:

- Examine joint angle performance measures such as measures of joint angular velocity and acceleration (Eman, 1999).
- Develop a new sequencing similarity measure in which the transition probabilities depend not only on the current state but on previous events as well.
- Investigate ways of quantifying human reliability and frequency of errors (Joice, 1998).
- Use distributions of performance measures (such as velocity) rather than summary measures such as mean velocity in the comparative analysis.
- Investigate the effects of fatigue loading on joint stress.

4.4 Future Studies

As mentioned previously, our long-term research goal is to develop validated simulations of surgical tasks based on motion analysis studies in the OR. Our intent is to use these controlled simulations as platforms for assessing surgeons’ performance and evaluating improved MIS instrumentation.

Several components of the developed methodology required further investigation to assess its overall reliability and clinical utility. The following questions remain unanswered:

- How reliable is our selection of performance measures in detecting differences between a surgeon in the OR and simulator and of surgeons of varying skill?
- Is the inter-procedural variability large compared with inter-surgeon differences or skill level differences?
• How much data do we need to detect measurable differences between various skilled surgeons? Is this amount feasible?
• What is the minimum number of performance variables needed to get a well-rounded measure of a surgeon’s skill level?
• Can we verify that time spent in a simulator translates into improved operating room performance and, if so, what is the “transfer effectiveness”?

The following sections outline relevant studies, which should be considered in achieving our long-term research goals and addressing the above questions.

4.4.1 Simulator Validation

The next logical step in this project is to begin the process of validating a surgical simulation. For this phase of the project, we propose to develop a simple desktop simulation (perhaps based on existing ones – e.g., Derossis 1998) which have tool and arm motions similar to those surgeons use in actual surgery. To validate the simulation, we must ensure that a surgeon treats the simulation in all relevant respects the same way they treat a live patient. The validation process requires, (1) showing reliability of our selected performance measures in the simulation, and (2) an iterative process of comparing performance measures between the OR and simulator to refine aspects of the simulation which will improve correspondence of these performance measures.

For the immediate future we suggest using the same data acquisition system and would expect to see some improvements in our marker trackability, as the simulator environment will be less complicated than the operating room and we will have better sight lines. We propose to use the quantitative measurements of surgical tool motions and surgeon posture from this OR study for our comparative evaluation of movements in surgical simulations. To measure the discrepancy between performance measures, we use the methodologies developed in this document.
The process of showing reliability of our performance measures is complex and requires us to assess how much difference in performance measure distributions is required to detect differences between OR and simulator tasks and between surgeons of varying skill levels. We expect to see some difference in distributions of performance measures collected in both cases. Since we intend to use the simulator as a tool for evaluating performance, we need to know how much difference we can detect at various performance levels. This requires measuring surgeons of different skill levels (described in the following section).

The iterative process of simulator validation requires making adjustments of the simulator so that it resembles, as closely as possible, the OR for a particular task. For example, consider joint posture - adjustment of table height and position with respect to the subject will help us to match joint angle distributions. Once we have established correspondence, we can introduce subjects of different skill levels and begin the process of showing performance measure reliability.

### 4.4.2 Surgical Skills Assessment

To establish reliability of our performance measure, we need to quantify the performance of surgeons of varying skill levels. For this assessment we suggest using a simulated setting which provides a more robust platform with less variability and influence from procedure-specific factors. It is also logistically much more feasible to get a wider range of subjects in the simulator, as it is often difficult to get surgeons with a wide range of skill levels in the OR.

To differentiate between surgeons with different skill levels, we need reliable and clinically meaningful performance measures which allow us to detect differences between them. The resolution of our ability to detect differences will depend on what differences in skill level we want to resolve. To address this issue we suggest collecting data from a large number of surgeons and develop a database of the recorded
performance measures. This database will allow us to provide context to our difference measurements and show reliability and feasibility. Feasibility will depend on the number of samples required for our desired resolution.

Finally we need to establish a ranking system for evaluating those measures which are most reliable at detecting differences in performance. In our global assessment of surgeon performance, measures with high reliability should be weighted higher than performance measures with lower reliability. This will allow us to consolidate our performance measures such that we can deliver useful clinical observations for making assessments and ways of suggesting improvements.
List of Terms

**Action States**

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HD</td>
<td>Hold</td>
</tr>
<tr>
<td>RH</td>
<td>Reach</td>
</tr>
<tr>
<td>RT</td>
<td>Retract</td>
</tr>
<tr>
<td>PH</td>
<td>Push</td>
</tr>
<tr>
<td>PL</td>
<td>Pull</td>
</tr>
<tr>
<td>TR</td>
<td>Translate</td>
</tr>
<tr>
<td>SW</td>
<td>Sweep</td>
</tr>
<tr>
<td>OR</td>
<td>Orient</td>
</tr>
<tr>
<td>CW</td>
<td>Clockwise-Orient</td>
</tr>
<tr>
<td>CCW</td>
<td>Counter Clockwise-Orient</td>
</tr>
</tbody>
</table>

**Ces**

- Cumulative Event Sequence
- Cumulative Event Sequence for all procedures
- Cumulative Event Sequence for all procedures excluding procedure i

**Cpd**

- Cumulative Probability Distribution
- Cumulative Probability Distribution for an individual procedure i
- Cumulative Probability Distribution for all procedures
- Cumulative Probability Distribution for all procedures excluding procedure i

**D**

- KS (D) statistic difference measure
- CPD of D's from resampling
- CPD of D's calculated from each procedure

\[
D^i_{measured} = D(CPD_{Ti}, CPD_i) \quad \text{where } i \text{ is the } i^{th} \text{ procedure}
\]

**Dreference**

- CPD of D's from resampling

\[
D_{reference} = RS_n^m(CPD_T)
\]

**Dref_meas**

- CPD of D's from resampling

\[
D_{ref, meas} = D(\delta_{reference}, \delta_{measured})
\]

**Ddelta**

- CPD of D's from resampling

\[
D_{delta} = RS_n^m(\delta_{reference})
\]

**Dref_meas**

- CPD of D's from resampling

\[
D_{ref, meas} = D(D_{reference}, D_{measured})
\]

**D_95**

- 95% confidence level of the bootstrap data

\[
D_{CPD, 95}^{95}\%
\]

**D_50**

- 50% confidence level of the bootstrap data

\[
D_{CPD, 50}^{95}\%
\]

**δ (delta)**

- Event sequencing difference measure
- CPD of δ's from resampling
- CPD of δ's calculated from each procedure

\[
\delta^i_{measured} = \delta(TPM_{Ti}, TPM_i) \quad \text{where } i \text{ is the } i^{th} \text{ procedure}
\]
\( \delta_{\text{reference}} = RS^n_n(\delta(\text{TPM}_T, \text{TPM}_{\text{random}})) \)

where \( \text{TPM}_{\text{random}} \) is the transition probability matrix for \( m \) randomly selected events in \( CES_T \)

\( ES_i \)  
Event Sequence for an individual procedure \( i \)

\( FFT \)  
Fast Fourier Transform

\( f_c \)  
Frequency cutoff

\( IRED \)  
Infra-red Emitting Diode

\( KS \)  
Kolmogorov-Smirnov

\( m \)  
resampled data sequence length

\( \text{MDMA} \text{Array} \)  
Multi Directional Marker Array

\( \text{MIS} \)  
Minimal Invasive Surgery

\( n \)  
number of events or samples

\( N \)  
number of data points

\( N_e \)  
Effective \( N \)

\( \text{NCL} \)  
Neuromotor Control Laboratory

\( \text{NWPSS} \)  
Normalize Weighted Postural Stress Score

\( p \)  
p-value

\( \text{Phases} \)  

\( \text{CDD} \)  
Cystic Duct Dissection

\( \text{GBD} \)  
Gall Bladder Dissection

\( \text{GBR} \)  
Gall Bladder Removal

\( r \)  
number of bootstrapping cycles

\( \text{RS} \)  
Resample

\( \text{RS}^m_r \)  
resample in length of \( m, r \) times

\( \text{RULA} \)  
Rapid Upper Limb Assessment

\( \text{Subtasks} \)  

\( \text{TT} \)  
Total Time

\( \text{AP} \)  
Approach

\( \text{TM} \)  
Tissue Manipulation

\( \text{CP} \)  
Clip Application

\( \text{WD} \)  
Withdrawal

\( \text{Surgical Tools} \)  

\( \text{Gr} \)  
Grasper

\( \text{S} \)  
Scissors

\( \text{Ir} \)  
Irrigation

\( \text{LGr} \)  
L-Grasper

\( \text{RGr} \)  
Ratcheted Grasper
\( Ca \) Camera
\( Sp \) Spatula
\( Cp \) Clipper

\( TPM \) Transition Probability Matrix
\( TPM_i \) Transition Probability Matrix of \( ES_i \)
\( TPM_T \) Transition Probability Matrix of \( CES_T \)
\( TPM_{Ti} \) Transition Probability Matrix of \( CES_{Ti} \)

\( \text{nVS}_p \) normalized Variance shift caused by patient variability
\( \text{VS}_p \) Variance Shift
Bibliography


Appendix A
OR Study Experimental Protocol and Data Acquisition Procedures

A.1 Experimental Protocol

Surgeon: Dr. Alex Nagy (VGH)
Researchers: Mr. Paul McBeth (UBC)
            Dr. Antony Hodgson (UBC)
Location: Vancouver Hospital
Procedure: MIS cholecystectomy

Study Protocol:

This is a protocol for the motion analysis study of Dr. Alex Nagy performing a MIS cholecystectomy. The surgeon is asked to scrub and enter the OR while the patient is being anesthetized. Before the operation begins, one researcher also scrubs, in order to attach the sterilized marker arrays to the surgeon’s hands, lower arms, torso, and surgical tools. The researcher remains scrubbed in to assist the surgeon and to be available to make any adjustments to the marker arrays or support cuffs. If the researcher is unable to scrub in, they instruct the OR nurse on the correct procedure for attaching the marker arrays from outside the sterile field.

Once the marker arrays are secured in place, another researcher outside the sterile field performs a test to ensure the motion analysis equipment is operational. Once the equipment is tested and confirmed operational, the researcher informs the surgeon that they are ready to begin recording. When the surgeon confirms he is also ready, the researcher begins tracking the markers arrays with the Polaris motion capture system and begins recording the operation using both the external video camera and laparoscopic camera. The equipment records for the duration of the initial open and subsequent laparoscopic portion of the surgery. When the laparoscopic portion of the surgery is completed, the surgeon informs the researcher and equipment is shut off.

The surgeon continues to wear the marker arrays once the surgery is completed in order to perform joint calibration. The researcher instructs the surgeon to perform a five-minute joint calibration procedure. Once the joint centre calibration procedure is complete, the surgeon is free to remote the marker arrays.

If the surgeon feels uncomfortable wearing the markers at any time during the surgery, they are free to stop the experiment and remove the markers. Approval for this experiment was granted though the University of British Columbia Clinical Research Ethics Board and the Vancouver Hospital. In additional all equipment has been approved by Sterile Supply and Biomedical Engineering Department.
A.2 Equipment List

The following outlines the hardware and software components required for experiments conducted in the OR:

**Hardware:**

1 - Panasonic video recorder
   - Panasonic AC adaptor
1 - 6’ extension cord
1 - Super VHS video cassette
1 - Portable computer desk
1 - Polaris Tool Interface Unit
1 - Polaris Position Sensor
1 - Serial port cable
1 - Polaris Position Sensor cable
1 - Polaris power cable
1 - Power bar
1 - Logitech web cam
1 - 6’ USB extension cord
2 - Tripods
1 - PC 800 MHz AMD Duron (tower, keyboard, mouse, monitor)
2 - Video tapes
1 - Digital camera
1 - Active marker array *
11 – Passive marker arrays and MDMArray’s *
   PSV1 – (3 reflective markers)
   PSV2 – (3 reflective markers, Velcro back)
   PSV3 – (3 reflective, 4 – 2mm hex bolts, 6 – 2mm nuts, 1” ACME clip)
   PSV4 – (3 reflective markers)
   PSV5 – (3 reflective, 4 – 2mm hex bolts, 6 – 2mm nuts, 1” ACME clip)
   PSV6 – (3 reflective markers)
   TOOL-1 (MDMArray) - (5 reflective markers, 3 – NDI mounting posts, 2 NCL mounting posts, 5 – 2mm hex bolts, 3 – 2mm nuts, 3 – 2mm washers, 1” ACME clip)
   TOOL-2 (MDMArray) - (5 reflective markers, 3 – NDI mounting posts, 2 NCL mounting posts, 5 – 2mm hex bolts, 3 – 2mm nuts, 3 – 2mm washers, 1” ACME clip)
   TOOL-3 (MDMArray) - (5 reflective markers, 3 – NDI mounting posts, 2 NCL mounting posts, 5 – 2mm hex bolts, 3 – 2mm nuts, 3 – 2mm washers, 1” ACME clip)
   TOOL-4 (MDMArray) - (5 reflective markers, 3 – NDI mounting posts, 2 NCL mounting posts, 5 – 2mm hex bolts, 3 – 2mm nuts, 3 – 2mm washers, 1” ACME clip)
TOOL-5 (MDMArray) - (5 reflective markers, 3 – NDI mounting posts, 2 NCL mounting posts, 5 – 2mm hex bolts, 3 – 2mm nuts, 3 – 2mm washers, 1” ACME clip)

TOOL-6 (MDMArray) - (5 reflective markers, 3 – NDI mounting posts, 2 NCL mounting posts, 5 – 2mm hex bolts, 3 – 2mm nuts, 3 – 2mm washers, 1” ACME clip)

2 – extra clips 1 – (1”ACME clip, 2 – 2mm hex bolts, 6 – 2mm nuts)

8 – ½” dia. 1” long surgical tubing *

4 – Stretchy Velcro straps *

1 – Chest marker attachment strap *

2 – Foam Velcro straps w/ mating end *

1 – 3/16” socket head w/ extension piece *

1 – 2mm Allen key *

* Equipment requiring sterilization

Software:

Windows 2000 or Window NT
Matlab 6.0 R12
Tera Term Pro V2.3
Logitech QuickCam V5.4.1
Matlab programs:
- PMCS.m
- Calibrate_joints.m
- OR_Tools_Calib.m
A.3 Marker Placement

Refer to the Figure A.1 for proper position and orientation of markers:

Actv1 – Torso
PSV2 – Right hand
PSV3 – Left tool
PSV4 – Distal Forearm
PSV6 – Proximal Forearm
Tool-1 – Spatula
Tool-2 – Scissors
Tool-3 – Clipper
Tool-4 – Grasper 1
Tool-5 – Grasper 2 (L-grasper)
Tool-6 – Grasper 3 (Ratcheted grasper)

*Figure A.1* – Marker placement to capture limb motion during surgery (L). Multi-Directional Marker Array on a dominant-hand laparoscopic tool handle (R).
A.4 OR Procedure

The following are the step-by-step procedures for motion capture data collection in the OR.

A.4.1 Pre-operative Set-up

Required Time: 30min – 1hr
Suggested start time: 0600 for a 0745 procedure

Set-up equipment as shown in Figure A.2.

Figure A.2 - Vancouver Hospital operating room equipment layout.
A.4.2 Software and Polaris Initialization

Required Time: 10min
Suggested start time: 0650 for a 0745 procedure

Start-up Procedure Teraterm

1) Run Teraterm
2) Turn on Polaris Tool Interface Unit (wait 20 sec for beep – \textit{RESETBE6F} will appear in the Teraterm command window)
3) Type: \textit{COMM 50000} – Reply: \textit{OKAYA896}
4) Teraterm window: Setup – Serial port… Change baud rate to 115200
5) Type: \textit{INIT_ enter} (note: _ means space bar) – Reply: \textit{OKAYA896}
6) Teraterm window: \textit{File – Exit}

Start-up Procedure Matlab (IR testing)

1) Start Matlab 6.0 R12
2) Change current directory: \textit{d:\Polaris}
3) Run IR\_test from Matlab command window
4) Check for any error messages
5) Identify external Infra-red sources (if required)
6) Shutdown Matlab

A.4.3 OR Data Collection Procedures

Required preparation time: 5 min
Required Time: Approximately 75 min
Suggested start time: 0745

1) Start video camcorder (focus camera on the surgeon upper limb and the surgical area
2) Start the MIS VCR to record the laparoscope
3) Start Matlab R12
4) Set Current Directory to: \textit{d:\Polaris\Polaris\_Interface\OR\_tracker}
5) In the Command Window type: \textit{PMCS}
6) Figure A.3 shown below will appear
7) Select the all radio buttons for all markers
8) STOP check equipment connections before initialization of Polaris
9) Press the \textit{Initialize Polaris} when ready.
10) Wait until all status bars are illuminated in yellow (Check for error messages in the Command Window).
11) In the Right Hand Tool Choice box select the appropriate tool used by the surgeon
12) Monitor the status bars and adjust equipment (ie. camera) if required. (Red status bar: missing marker, Yellow status bar: loading tool identification files, Green status bar: tracked marker)

13) At the end of the procedure press Stop Tracking → followed by Shut-down Polaris (Do not do this in the opposite direction).

Figure A.3 – Custom designed data acquisition software (PMCS.m)

### A.4.4 Joint Centre Calibration Data Collection Procedure

Required preparation time: 1 min  
Required Time: 5 min  
Suggested start time: immediately post-operative

1) Set Current Directory to: d:\Polaris\Calibration\Joint_Calib_Data_Collection  
2) In the Command Window type: Calibrate_joints  
3) Follow instructions provided on the Matlab Command Window (Note: each sequence tracks 200 samples)
4) Joint centre calibration  
   - Right shoulder centre calibration – Shoulder abduction / adduction - Subject must hold elbow locked while swinging the arm in an approximately cross shaped arc incorporating shoulder abduction / adduction, flexion / extension, and internal / external rotation.
- Right elbow centre calibration – Elbow flexion / extension - Subject must hold upper arm vertically downwards while swinging the forearm about the elbow in flexion and extension.
- Right forearm axis calibration – Forearm pronation / supination - Subject must hold upper arm vertically downwards and elbow at 90° while rotating the forearm in pronation / supination motion.
- Right wrist centre calibration
  - Wrist flexion / extension – Subject moves wrist in flexion and extension.
  - Wrist radial / ulnar deviation – Subject moves wrist in radial and ulnar direction.
  - Wrist spherical – Subject rotates wrist in a spherical motion.
- Neutral Pose – limbs positional with zero reference position – Shoulder down, elbow flexed at 90°, straight wrist and hand.

5) After each calibration step a plot will appear. Data points should be in definable paths. If there are stray data points redo the calibration for that step. Follow direction in Matlab Command Window.

6) Before the surgeon removes the markers check for the following files in directory: 
   \texttt{d:\Polaris\Data\storage\Joint\calibration:}
   neutral\_r.mat
   re\_fe\_calib.mat
   re\_ps\_calib.mat
   rs\_calib.mat
   rw\_calib.mat
   rw\_fe\_calib.mat
   rw\_ru\_calib.mat

7) Remove markers from surgeon

\textbf{A.4.5 Tool Calibration Data Collection Procedure}

Required preparation time: 10 min
Required Time: 30 min
Suggested start time: post-operative

1) Start Matlab 6.0 R 12
2) Change current directory: \texttt{d:\Polaris\OR\Tool\calibration}
3) Run: Tool\_calibration
4) Place the tip of the tool at a stationary location and rotate the tool handle in a spherical motion about this point
5) Select: PSV1 – extra tool
   PSV3 – Left hand tool
   Tool-1 – Spatula
   Tool-2 – Scissors
   Tool-3 – Clipper
   Tool-4 – Grasper 1
   Tool-5 – Grasper 2
   Tool-6 – Grasper 3
6) Refer to Matlab Command Window for tracking update. Note the sequence will track 150 times.
7) Check plot for stray data points
8) Remove marker from each tool
9) Submit for markers and tools for sterilization

A.4.6 Clean-up Equipment / Back-up data

Required Time: 15 min
1) Account for all equipment (including marker array sub components) outlined in Section A.2.
2) Submit any contaminated equipment for sterilization.
3) Back-up data on hard-drive.

A.5 Contacts

See Dr. Antony Hodgson for list of hospital contacts.
Appendix B
Measurement Equipment and Data Acquisition Protocols for use in the Operating Room

B.1. Polaris System

In our experiments we used a Polaris Hybrid Optical Tracking System (Northern Digital Inc. (NDI), Waterloo, Ontario, Canada) as shown in Figure B.1 The system tracks the three-dimensional position of both passive retro-reflective markers and active infra-red light emitting diodes (LEDs). The positional accuracy of each marker is 0.35 mm. Polaris dynamically tracks and calculates the location (X,Y,Z spatial coordinates) and orientation (roll, pitch, yaw angles) for each marker array. A marker array is defined by a minimum of three markers on a rigid frame. The system is capable of tracking three active arrays and nine passive marker arrays.

![Figure B.1 - Polaris Hybrid Optical Tracking System (Northern Digital Inc., Waterloo, Ontario, Canada). System components: Position Sensor (top), Tool Interface Unit (middle), Marker Arrays (both active and passive markers) (bottom).](image)
We used Polaris to track the position of one active array and nine passive arrays at 20 Hz with occasional missing data points. The active array and three passive arrays were used to collect ergonomic postural data. The remaining six passive arrays were used in tool tracking; one passive array on the left hand tool (non-dominant hand) and five passive arrays on the right tool (dominant hand). See Figure B.2.

![Image of markers arrays used to track position of the upper limb (left), non-dominant hand tool (top right), and dominant hand tool (bottom right).]

**Figure B.2** – Markers arrays used to track position of the upper limb (left), non-dominant hand tool (top right), and dominant hand tool (bottom right).

### B.1.1 Planar Marker Array Design

Passive and active marker arrays are made from 1/8” plate aluminum machined into an ‘L-shaped’ based as shown in Figure B.3. For passive arrays, three retro-reflective markers are mounted to the array plate with stainless steel machine screws and for active arrays, three ¼” active infra-red LEDs are mounted to the array plate with epoxy.

This planar marker array design is limited in its trackable range of movement as only one plane is visible to the camera. If the plane of the array is tilted more than ~80° from the plane of the camera then trackability is lost. The total angular range of movement is ~160° for planer marker arrays.
B.1.2 Multi-directional Marker Array Design (MDMArray)

In minimally invasive surgery the surgical tool is constrained to three rotational degrees of freedom (Roll, Pitch, Yaw), and 1 translation degree of freedom (Plunge). The limited trackability of planar marker arrays prevents continuous tracking of rotating surgical tools where the face of an array can be rotated out of line-of-sight. There is a need for a multi-directional marker array.

A new marker array was developed to limit occlusions resulting from the wide range on surgical tool movements. The Multi-Directional Marker Array (MDMArray) consists of multiple faces, which can be tracked simultaneously by Polaris. The MDMArray is constrained to having geometrically unique faces (as outlined by Polaris) and must be lightweight, compact, sterilizable, easily attached and released from the surgical tools and well accepted by surgeons.

The resulting MDMArray, as shown in Figure B.4, measures 72 mm in diameter and 58 mm high and weights 26 grams and is built from 6061 aluminum alloy. The aluminum was polished and finished with a black anodized surface finish to minimize excess infrared light reflections. It has five detachable retro-reflective markers, which form five geometrically unique faces. Three of the detachable markers are attached to the base using NDI mounting posts. The remaining two detachable markers are attached to custom designed mountain posts. The MDMArray is equipped with a ¾” metal binder clip allowing easy attachments and release to the shaft of the instrument. The resulting MDMArray was well accepted by the surgeon and hospital staff.

The range of marker trackability has been enhanced. The MDMArray can be rotated about its longitudinal axis with ~350° of trackability. Normal to the longitudinal axis the trackability is ~160°.
Figure B.4 – Multi-Directional-Marker-Array (MDMArray) – A multi faced marker array used to track the position of a laparoscopic surgical tool (left). The MDMA attached to a laparoscopic surgical tool (right).

A quantitative evaluation of the MDMAarray was made using a bench top simulator. The trackability of the MDMAarray was calculated and compared with a planar marker array for 30 simulated manipulation tasks using a standard laparoscopic tool. An average improvement in trackability of 159% (SD: 88%) was found. See Figure B.5.

Figure B.5 – A percent trackability evaluation of MDMAarray and Planar marker arrays attached to a standard laparoscopic tool for 30 simulated manipulation tasks.
B.2. Video System

A video system was used in conjunction with the Polaris system in order to correlate information about the actual surgical procedures to corresponding raw motion data tracked by the optoelectronic system. Video images of the surgery were collected using both a laparoscope and an external camera aimed at the surgeon. The images from these two sources were time stamped and recorded onto a standard VHS tape. The laparoscope system was comprised of a standard 10mm - 30° surgical laparoscope, camera and illuminator (Stryker Endoscopy). These components are arranged and connected in the OR as illustrated in Figure B.6.

![Figure B.6 - Data collection components – Polaris Hybrid Optical Tracking System and Video Recording Equipment.](image)

B.3. Data Acquisition Protocol

One expert surgeon was evaluated in seven clinical laparoscopic cholecystectomies over a period of four months at Vancouver Hospital (Vancouver, British Columbia). Posture data and tool tip trajectories were measured with optoelectronic motion analysis equipment. Sterilized planer marker arrays are attached to the surgeon’s torso and dominant arm (on the proximal and distal forearm and on the hand), as shown in Figure B.2. The torso array was made of active markers; all others were passive. Marker arrays were secured to the surgeon using elastic and Velcro® harnesses. Individual MDMArray’s were attached to each surgical tool (scissors, clipper, spatula, tissue grasper, ratcheted grasper, locking grasper). The Polaris system was connected to a personal computer (800 MHz AMD Duron) running custom designed Matlab data acquisition software. Approval for this experiment was granted through the University of British Columbia Clinical Research Ethics Board and the Vancouver Hospital. Additional details in Appendix A.
Appendix C
Signal Processing and Data Formatting for Optoelectronic Marker Array data from the OR

C.1. Signal Processing

For the positional tracking of surgical tools and limbs in space an optoelectronic system was chosen because of its sterilizability, sampling properties and its availability despite its inherent line-of-sight limitations. This section describes properties of the signals returned from Polaris and the methods used to deal with missing data cause by occlusions.

C.1.1 Missing Data

The missing data samples are a result of occlusions, internal localization errors or additional processing time required by Polaris. Missing data resulting from occlusions is a serious problem in tracking the surgical instruments using optoelectronic systems. Occlusions occur when the line-of-sight between the marker and the camera becomes obscured. Obstructions may include the surgeon’s hands or surgical instrumentation. During surgery an attempt to minimize occlusions was made by carefully positioning the optoelectronic camera to maximize workspace trackability and by attaching markers in the most visible position and orientation. Custom designed tool marker arrays where developed to help deal with line-of-sight limitations as discussed in Appendix B.

Internal localization errors or ‘flicker’ occurs when Polaris is unable to find a marker within 0.5 mm of its expected location. These errors are the result of poor marker calibration constants derived from marker array geometry or the inability of Polaris to distinguish circular characteristics of each reflective sphere. See Figure C.1. As a result a marker array which is clearly in the field of view may temporarily not register with Polaris.

![Figure C.1 - NDI Retroreflective Marker Visibility ~164° hemispheric range of view. The internal localization of the marker is based on the centroid. Side view (left) has a shifted centroid which introduces localization errors. Top view (right) has not shifted centroid.](image)

*Figure C.1 - NDI Retroreflective Marker Visibility ~164° hemispheric range of view. The internal localization of the marker is based on the centroid. Side view (left) has a shifted centroid which introduces localization errors. Top view (right) has not shifted centroid.*
Unequally spaced sampled data is the result of real-time calculations required by Polaris. The real-time fitting calculations of the Polaris system often require one or two extra (1/60 s) clock frames. The frequency of these additional computation periods depends on the number of passive markers tracked simultaneously (more markers, more computation).

C.1.2 Signal Properties

The signal returned by Polaris is the position (X,Y,Z spatial coordinates) and orientation (q1, q2, q3, q4 quaternions) for each marker array. Analysis of the spectral content of each signal found each having approximately the same cutoff frequency and noise signal properties. The Fast Fourier Transform (FFT) and residual power plots on a sample signal are shown in Figure C.2. The underlying frequency content ($f_c$) is $\sim 4$ Hz.

![Figure C.2 - Spectral signal properties. Signal power as a function of signal frequency (left). Residual signal power as a function of signal frequency. $f_c = \sim 4$Hz (right).](image)

C.1.3 Filtering and Resampling

The non-equally sampled nature of the data limits our ability to filter noise and interpolate missing points. Traditional frequency based techniques rely on continuously sampled data at a fixed sample frequencies. Resampling non-equally spaced data at a fixed rate using interpolation techniques prior to filtering is not advised because of signal noise enhancement.

Spline fitting techniques are often used on unequally spaced data (Farin 1988). Spline coefficients are calculated and based on sampled data points. A smoothing parameter (S) was used to describe the smoothness of a spline and was chosen to reduce measurement noise. A Generalized Cross Validation (GCV) technique was used to optimally find a smoothing parameter for fitting the spline to the data set. The GCV method used in our experiment was originally developed by Woltring and has been used extensively in the biomechanics literature (Woltring 1986).
The GCV technique was used to optimally find coefficients of cubic and quintic splines fitting position data. A cubic spline was used to fit position data acquired at the tool handle (MDMArray attachment point). Using a homogeneous transformation matrix (found during tool calibration) the fitted position data of the tool handle is used to locate the tool tip. A quintic spline is fitted to the tool tip trajectory and used to calculate higher order derivatives. Quintic splines were chosen over the more commonly used cubic splines, for its higher order derivative continuity properties. The drawbacks of higher order splines are increased computation time and artifacts in positional interpolation (Curtis 1989). In certain applications quintic splines show oscillatory behavior in the interpolation of position data (Curtis 1989). Since our data at the tool tip does not require interpolation no oscillatory behavior was observed in our filtered tool tip position data. Additional details regarding the tool tip calibration process are found in Section C.2.2.

C.1.4 Interpolation / Extrapolation Errors

Interpolation and extrapolation were used to deal with missing points outlined in Section C.1.1. The error associated with interpolating across varying gap sizes was calculated. Five 20 sec samples of real data with no occlusions were used to calculate average and maximum RMS error of missing data points. The data set collect was real OR data and based on a sample of mixed movement types (rapid movements and slow movements).

The original test signal was smoothed using GCV filter. The coefficients from the spline fitting procedure are used to calculated velocity, acceleration, and jerk. A known sample point was removed from the original data set a specified time instant. The missing data set was smoothed and the new position, velocity, acceleration and jerk were calculated at the given specified time. The differences between the original and interpolated values are used in calculation of RMS error. The process was continued with larger gap sizes. The RMS mean difference and max difference was found for each gap size up to 15 samples. In the segments interpolated all possible data points on each side of the missing gap were used in the calculation. A minimum of two data points are required on each side of a missing data gap. Figure C.3 (upper left) shows a graphical representation of missing data interpolation for a segment of data.

The purpose of incrementing the missing segment along the total segment is to investigate the effect of gap size at different characterized movements. Movements can be fast or slow (high and low frequency). For slow movements the interpolation gap size is large and for fast movements the interpolation size is small. Because of the random nature of the position signal it is difficult to predict the frequency content of the data to vary the gap size. A gap size was chosen to be a compromise between larger errors in fast movements and smaller gap size interpolation in slow movements.

An RMS error of 1mm was chosen as an acceptable error which suggests a maximum gap size of 0.5 sec (10 samples) can be interpolated. A gap size of 10 samples allows a resolution in velocity, acceleration, and jerk of 22 m/s, 280 m/s², 4700 m/s³ respectively. To remain within the 1 mm RMS error boundary only two extrapolated data points are allowed. See Figure C.3.
Figure C.3 — Missing data interpolation and extrapolation. Graphical representation of missing data interpolation (top left). RMS interpolation error associated with measuring position and calculating velocity, acceleration, and jerk as a function of gap size (top right – bottom left). RMS extrapolation error associated with measuring position (bottom right).

C.1.5 Velocity, Acceleration, and Jerk Signal Properties

Calculating higher order derivatives of position data such as velocity, acceleration, and jerk leads to noise amplification and signal distortion. The frequency content of interest in the position signal is less than 4 Hz. This is considerably below the Nyquist frequency of 10 Hz which is required to prevent aliasing when sampling a 20 Hz. The sampling rate is appropriate for our analysis of position data. The accumulated power of the position, velocity, acceleration, and jerk signals suggests each signal is appropriate for our analysis as each signal contains at least 93% of its total power at 4 Hz. See Figure 4.
Figure C.4 – Accumulated normalized power for position, velocity, acceleration, and jerk signals.

C.2 Data Processing and Formatting

The output signal returned from Polaris is the position and orientation of each marker array. A series of formatting steps are required for identifying useful ergonomic joint angle data and tool tip data. The following sections describe the details and procedures of data formatting.

C.2.1 Ergonomic Data

Planar marker arrays are attached to the surgeon’s torso and dominant arm (on the proximal and distal forearm and on the hand). The position data from these marker arrays was used in combination with the post-operative calibration data to find joint angles. Figure C.5 illustrates the steps required to find joint angle data. Kinematic model details are outlined in Person’s thesis (Person 2000).
Figure C.5 – Ergonomic marker data processing and formatting flow chart.

C.2.2 Tool tip Data

Individual MDMArray’s were attached to each surgical tool (scissors, clipper, spatula, tissue grasper, ratcheted grasper, L-grasper). The MDMArray has five faces each of which are individually tracked by Polaris. The position data from these faces was used in combination with the post-operative (sphere fitting) calibration data, and filtering algorithms to find the position, velocity, acceleration and jerk of the tool tip. Figure C.6 illustrates the steps required to find tool tip data.
Five geometrically unique faces tracked on the MDMArray
FACE1, FACE2, ..., FACE5

Apply transform to tool tip
Tool tip transformation based on postoperative tool calibration using sphere-fitting techniques

Calculate common tool handle reference frame
Tool_ref_transforms.m

Optimally fit a cubic spline using GCV algorithm
Tool_handle_spline.m

Interpolate / Resample
Tool_handle_resample.m

Refit a quintic spline using GCV algorithm
Tool_tip_spline.m

Use spline coefficients to calculate position, velocity, acceleration, and jerk at the tool tip

**Figure C.6** - MDMArray marker data processing and formatting flow chart.

**C.2.3 Video Data**

The recorded video images of the surgery from both the laparoscope and the external camera aimed at the surgeon require synchronization. The images from these two sources were synchronized, time stamped and recorded together onto a standard VHS tape using a SONY SVO-9500MD SVHS.
C.3 Performance Measure Extraction

We developed custom design software, written in Matlab, for managing time durations, tool tip kinematics and joint angle data. The software is useful for visualization purposes and extracting useful performance measures. The graphical user interface (GUI) is easy to use and is equipped a comment window and a help file for assistance. See Figure C.7.

Figure C.7 – Custom designed trajectory-viewing program with performance measure displays. (SVP.m)
Appendix D
Operational Definitions

D.1 Hierarchical Decomposition Operational Definitions

To characterize and assess the repeatability of our chosen performance measures, we created a hierarchical decomposition describing the procedure in terms of surgical phases and stages, tool tasks and subtasks, and fundamental tool actions. This five-level hierarchical decomposition provides the foundation for a quantitative analysis of surgeon performance. This section provides a summary of the notation and terminology used in the hierarchical decomposition of a laparoscopic cholecystectomy. This framework is general enough to be expanded to apply to other surgical procedures. Figure D.1 shows the five levels of the hierarchical framework.

Figure D.1 – Five levels of the hierarchical decomposition.

D.1.1 Phase Level

Phases are the fundamental levels of a procedure forming the backbone and the foundation for further decomposition. A laparoscopic cholecystectomy procedure is divided into five distinct phases as shown in Figure D.2. Each phase has particular global goals and must be achieved before advancing to the next phase. Definitions of each phase are summarized in Table D.1.

Figure D.2 – Phase levels of a laparoscopic cholecystectomy.
### Table D.1 – Hierarchal phase level definitions.

<table>
<thead>
<tr>
<th>Phase Name</th>
<th>Abbrev.</th>
<th>Definition</th>
<th>Start</th>
<th>Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trocar Preparation</td>
<td>TP</td>
<td>Time spent preparing and inserting surgical tools in the abdominal wall of the patient</td>
<td>Administration of local anaesthetic to the naval (where the first trocar is place for the camera).</td>
<td>First indication the tools are used to manipulate tissue.</td>
</tr>
<tr>
<td>Cystic Duct Dissection</td>
<td>CDD</td>
<td>Time spent dissecting the omental foramen and tissue surrounding the cystic duct and artery.</td>
<td>First indication the tools being used to manipulate tissue.</td>
<td>Separation of all connecting vessels to the gall bladder.</td>
</tr>
<tr>
<td>Gall Bladder Dissection</td>
<td>GBD</td>
<td>Time spent dissecting the gallbladder from the gallbladder fossa on the visceral surface of the liver.</td>
<td>Following the CDD with most often the spatula used to start dissection.</td>
<td>Separation of the gall bladder from the liver.</td>
</tr>
<tr>
<td>Gall Bladder Removal</td>
<td>GBR</td>
<td>Time spent placing the gallbladder in a bag and removing it from the abdominal cavity.</td>
<td>Separation of the gall bladder from the liver.</td>
<td>Removal of the gallbladder from the abdominal cavity.</td>
</tr>
<tr>
<td>Closure</td>
<td>CL</td>
<td>Time spent removing trocars and suturing incisions.</td>
<td>Removal of the gallbladder from the abdominal cavity.</td>
<td>Final gauze or medical tape strips are placed.</td>
</tr>
</tbody>
</table>

Note: all phase level definitions based on video observation

### D.1.2 Stage Level

The phase levels of a procedure are further divided into stages as shown in Figure D.3. Each stage is characterized by its own end goal, however each defined stage within a phase is not required for successful completion of that phase. For example the CDD phase may be successfully completed without having to control bleeding if there was no bleeding to control. The start and ending of each stage is defined in terms of context of the tools being used within each stage. Definitions of each phase are summarized in Table D.2.
**Table D.2 – Hierarchal stage level definitions.**

<table>
<thead>
<tr>
<th>Phase</th>
<th>Stage</th>
<th>Tool used</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDD</td>
<td>Exploration (Required)</td>
<td>graspers, camera</td>
<td>Assess liver state, identify anatomical variations and locate various anatomical structures.</td>
</tr>
<tr>
<td></td>
<td>Isolate cystic duct and artery (Required)</td>
<td>graspers, L-graspers, irrigation, spatula</td>
<td>Locate and isolate cystic duct and cystic artery and to remove any adhesions.</td>
</tr>
<tr>
<td></td>
<td>Separate cystic duct and artery (Required)</td>
<td>clippers, scissors, spatula</td>
<td>Clip and cut major connecting vessels such as the cystic duct and the cystic artery.</td>
</tr>
<tr>
<td></td>
<td>Control bleeding (Optional if required)</td>
<td>clippers, graspers, spatula</td>
<td>Control any significant bleeding.</td>
</tr>
<tr>
<td>GBD</td>
<td>Isolate gallbladder from liver (Required)</td>
<td>graspers, irrigation, spatula</td>
<td>Separate the gallbladder from the liver.</td>
</tr>
<tr>
<td></td>
<td>Control bleeding (Optional if required)</td>
<td>clippers, graspers, spatula</td>
<td>Control any significant bleeding.</td>
</tr>
<tr>
<td>GBR</td>
<td>Clean-up (Optional if required)</td>
<td>irrigation</td>
<td>Remove any excess debris and to assess the state of any bleeding inside the abdominal cavity.</td>
</tr>
<tr>
<td></td>
<td>Bag gallbladder (Required)</td>
<td>ratcheted graspers, graspers</td>
<td>Insert the gallbladder into a baggie.</td>
</tr>
<tr>
<td></td>
<td>Remove gallbladder (Required)</td>
<td>ratcheted graspers</td>
<td>Remove the gallbladder from the abdominal cavity.</td>
</tr>
</tbody>
</table>

*Note: all stage level definitions based on video observation*
D.1.3 Task Level

A task is a set of movements performed with a single tool to achieve a desired effect. A number of tasks or tool sequences may be required to successfully complete a stage of a procedure. A task segment is defined from the time the tool tip is place in the distal end of the trocar until the tool is pulled out through the same trocar. The initial velocity is approximately zero and the movement of the tool into the trocar should be continuous without external interruptions. The withdrawal ends with the tool tip exiting the distal end of the trocar (independent of tool tip velocity). We selected the clip application as an example task, which is performed at least 6 times per procedure, to form the bases of the remaining decomposition analysis.

D.1.4 Subtask level

The subtask level describes how the surgical tool is moving inside the patient. Figure D.4 shows the subtask levels required for application of a clip and Table D.3 outlines definitions of each subtask.

![Subtask level diagram for clip application.](image)

**Figure D.4** - Subtask level diagram for clip application.

<table>
<thead>
<tr>
<th>Table D.3 – Hierarchal subtask level definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtask Name</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>FSM* - Approach</td>
</tr>
<tr>
<td>Tissue Manipulation</td>
</tr>
<tr>
<td>Clip Application</td>
</tr>
<tr>
<td>FSM* - Withdrawal</td>
</tr>
</tbody>
</table>

Note: all subtask level definitions based on video observation

* Free space movement (FSM) – when the tool is moving in freespace (no tissue manipulation).
D.1.5 Action level

Actions stand as the building blocks of all manipulation tasks and are based on 12 distinct types of tool movements. Of the 12 actions, three are based solely on video observations, two are based solely on kinematic data and the remaining seven are based on a combination of both kinematic and video observations. The kinematic and video data used to distinguish actions often suggest membership in more than one action state. As a result it is possible for tool movements to be in some combination or a collection of actions. In total there are 72 feasible combinations of these 12 basic actions.

Using the kinematic and video data the membership value of the 12 actions at each sample point was evaluated. A membership value ranges from zero to one; zero representing no membership and one representing complete membership. Table D.4 outlines the characteristic details defining each action state according to the reference frame defined in Figure D.5 (left).

<table>
<thead>
<tr>
<th>Action</th>
<th>State Abbrev</th>
<th>Jaw Position</th>
<th>Tissue Contact</th>
<th>Tool tip path as per Fig D.5</th>
<th>State Assessment</th>
<th>Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle</td>
<td>ID</td>
<td>Either</td>
<td>either</td>
<td>none</td>
<td>0.15 sec</td>
<td>abs vel. &lt; 2 mm/s for 0.15 sec</td>
</tr>
<tr>
<td>Hold</td>
<td>HD</td>
<td>Closed</td>
<td>yes</td>
<td>any</td>
<td>video</td>
<td></td>
</tr>
<tr>
<td>Reach</td>
<td>RH</td>
<td>Either</td>
<td>no</td>
<td>positive Z</td>
<td>both</td>
<td></td>
</tr>
<tr>
<td>Retract</td>
<td>RT</td>
<td>Either</td>
<td>no</td>
<td>negative Z</td>
<td>both</td>
<td></td>
</tr>
<tr>
<td>Push</td>
<td>PH</td>
<td>Either</td>
<td>yes</td>
<td>positive Z</td>
<td>both</td>
<td></td>
</tr>
<tr>
<td>Pull</td>
<td>PL</td>
<td>Closed</td>
<td>yes</td>
<td>negative Z</td>
<td>both</td>
<td></td>
</tr>
<tr>
<td>Translate</td>
<td>TR</td>
<td>Either</td>
<td>no</td>
<td>XY - plane</td>
<td>both</td>
<td></td>
</tr>
<tr>
<td>Sweep</td>
<td>SW</td>
<td>Either</td>
<td>yes</td>
<td>XY - plane</td>
<td>both</td>
<td></td>
</tr>
<tr>
<td>Orient</td>
<td>OR</td>
<td>Either</td>
<td>either</td>
<td>about Z-axis</td>
<td>kinematic &gt; 20 deg/s</td>
<td></td>
</tr>
<tr>
<td>CW</td>
<td>CW</td>
<td>Either</td>
<td>either</td>
<td>positive rot.</td>
<td>kinematic &gt; 20 deg/s</td>
<td></td>
</tr>
<tr>
<td>CCW</td>
<td>CCW</td>
<td>Either</td>
<td>either</td>
<td>negative rot.</td>
<td>kinematic &gt; 20 deg/s</td>
<td></td>
</tr>
<tr>
<td>Grasp</td>
<td>GR</td>
<td>Closed</td>
<td>either</td>
<td>any</td>
<td>video</td>
<td></td>
</tr>
<tr>
<td>Release</td>
<td>RL</td>
<td>Open</td>
<td>yes</td>
<td>any</td>
<td>video</td>
<td></td>
</tr>
<tr>
<td>Spread</td>
<td>SP</td>
<td>Open</td>
<td>either</td>
<td>any</td>
<td>video</td>
<td></td>
</tr>
</tbody>
</table>
D.1.5.1 Action Level Membership Functions

The kinematic definitions of the reach, retract, and translate actions are the same as the push, pull, and sweep actions respectively; the only difference being tissue contact (established by video observation). For this discussion consider these actions as forward, backward, and sideways. The projection of the tool tip path vector on the tool axis vector at each sample forms the bases for establishing forward, backward, and sideways membership (See Figure D.5(right)). The projected vectors and angle between vectors is calculated at each sample interval. The angle between vectors is used to establish the direction of tool movement i.e. forward (0°-90°) or backward (90°-180°). A normalization of the projected vectors is used to calculate the membership of forward and backward, or sideways movement. The membership functions (MF) for the axial (forward / backward) and normal (sideways) directions is calculated as follows:

\[
\text{MF}_{\text{axial}} = \frac{\left(\text{proj}_{T_{r}}^T\right)^2}{\left(\text{proj}_{T_{r}}^T\right)^2 + \left(\text{proj}_{T_{n}}^T\right)^2}, \quad \text{MF}_{\text{normal}} = \frac{\left(\text{proj}_{T_{n}}^T\right)^2}{\left(\text{proj}_{T_{r}}^T\right)^2 + \left(\text{proj}_{T_{n}}^T\right)^2} \quad [D.1]
\]

Transitions between states are often rapid and membership duration is short-lived making it difficult to tell when the transitions started and ended. Figure D.6 shows sample membership functions for selected actions for a clipping task. For our performance measure analysis we considered mean membership values however, we did make an attempt at isolating tool tip actions states by establishing a series of arbitrary rules and cut-offs and used this data in our sequencing analysis presented in Appendix F.
D.2 Performance Measures Definitions

D.2.1 Kinematic Performance Measures

Table D.5 outlines the kinematic performance measure definitions used for quantifying performance at the tasks and subtask levels of the hierarchal decomposition.

**Table D.5 – Time and kinematic performance measure definitions**

<table>
<thead>
<tr>
<th>Kinematic Measure</th>
<th>Performance measure</th>
<th>Kinematic Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Start / stop definitions outlined in Section D.1</td>
<td>$T = t_{stop} - t_{start}$</td>
</tr>
<tr>
<td>Distance</td>
<td>Path distance</td>
<td>$d = \sum_{i=\text{start}}^{\text{stop}} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}$</td>
</tr>
<tr>
<td>Velocity</td>
<td>Minimum</td>
<td>$v_i = \sqrt{\dot{x}_i^2 + \dot{y}_i^2 + \dot{z}_i^2}$</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td>Acceleration</td>
<td>Minimum</td>
<td>$a_i = \sqrt{\ddot{x}_i^2 + \ddot{y}_i^2 + \ddot{z}_i^2}$</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td></td>
</tr>
<tr>
<td>Jerk</td>
<td>Jerk cost (Smoothness) (Nelson, 1983)</td>
<td>$J = \frac{1}{2} \int_0^T \dddot{x}^2(t) dt$</td>
</tr>
<tr>
<td>Trajectory deviation</td>
<td>RMS straight line deviation</td>
<td>RMS error between measured trajectory and a straight-line trajectory</td>
</tr>
</tbody>
</table>
D.2.2 Sequence Similarity Measure

Transition state diagrams were used to represent event sequences and describe the probability of transitions between different states. The transition probabilities are calculated by dividing the number of transitions between one state and a second state by the total number of transitions from the first state to all other possible states. The transition probabilities of the state diagram are reduced to a Transition Probability Matrix (TPM). Each cell of the transition probability matrix (TPM) represents the transition probability from one event to another (event i to event j). An example of a state transition diagram and TPM are shown in Figure D.7

**Figure D.7** – Example of an event sequence, state transition diagram, and probability transition matrix.

To establish the confidence interval for the transition probabilities we use the following equation (Walpole, 1978):

\[
\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p} \hat{q}}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p} \hat{q}}{n}}
\]  

[D.2]

Where \( \hat{p} \) is the estimated transition probability, \( n \) is the number of transitions used to calculate the transition probability, and \( q=1-p \).
Appendix E
Preliminary Selection of Performance Measures and Development of Comparative and Reliability Analysis Techniques

E.1 Development of Performance Measures

During the course of our data analysis we considered a number of techniques to identify and extract meaningful and reliable performance measures. In addition to our selected performance measures we considered the following analysis techniques:

- Spectral analysis using Fast Fourier Transforms (FFT) and short-time Fourier transforms
- Wavelets
- Trajectory template matching

The spectral and wavelet analysis techniques were considered of limited utility in our application. Using these techniques on our kinematic data we found no obvious extractable performance parameters. Template matching may be useful in identifying performance measures related to the tool tip trajectories, however due to limited time we were unable to implement this technique.

E.2 Performance Measure Comparative Analysis Using the K-S Statistic

To compare distributions of performance measures in different contexts, we use the Kolmogorov-Smirnov (K-S) statistic. The K-S statistics is useful because it is nondimensional, and makes no a priori assumptions about the nature of the statistical distributions of the data. In addition it considers the entire distribution rather than other statistics, which only test for differences in mean ($t$-test) or variance ($f$-test) (Von Mises, 1964). The K-S statistic is commonly used to test the null hypothesis that two sets of data were drawn from the same underlying distribution. This is helpful for testing the significance of procedural specific variability discussed in Section E.2.3.

E.2.1 Establishing Confidence Bounds

In our applications the K-S statistic is computed based on two finite sets of measures, the value we find is itself an estimate of the true difference between the two sets. To assign a confidence interval to our estimate we use a bootstrapping approach. By resampling the two finite sets of measures and evaluating the new $D$ value (for resamples $\gg 0$) we establish a distribution of $D$ values as shown in Figure E.1.
**Figure E.1** – Resampling distributions of two finite sets to establish confidence bounds on our estimate for $D$. The right illustration depicts how we establish the 5% and 95% confidence bounds on our estimate for $D$.

### E.2.2 Correction of $D$ for Sample Size

Due to the limited size of our data our estimated values for $D$ are consistently greater than the actual value. This is also the case when the difference in means is low. For our purposes we correct the bias in the $D$ estimate by introducing the data vector length. Through numerical experimentation conducted by Hodgson (2002), we found the following function gives an approximation of $D$ (our estimate of $D$):

$$
\bar{D} = D + \bar{D}_0(N) \times \frac{0.7742D + 0.0078}{1 - D(1 - 1/\bar{D}_0(N))}
$$  \[E.1\]

Where $\ln \bar{D}_0(N) = k_1 + k_2 \ln N$ and $k_1 = 0.0261$, $k_2 = -0.471$. $N$ is the number of elements in the data vector. For cases where the data vectors are of unequal length we use the effective length $N_e$ given by:

$$
N_e = \frac{N_1N_2}{N_1 + N_2}
$$  \[E.2\]

Given $\bar{D}$ and $N$, we solve the quadratic equation in $D$ from equation E.1 to obtain an estimate of $D$. 
E.2.3 Reliability / Patient Variability

As part of our analyses to show reliability of our performance measures we attempt to quantify the effect of procedural variability. This section is intended to outline some of the additional computational details omitted from the methods section of Chapter 2. The goal is to measure the difference between two distributions which represent: (1) the range of differences we would expect if we measured our surgeon on a standard (ideal) patient ($CPD(D_{\text{reference}})$) and (2) the range of differences we would expect to get if we measured this surgeon in the OR on an arbitrary patient ($CPD(D_{\text{measured}})$). The difference between these two distributions reflects the amount of procedural variability. The example provided is the distribution of clipping task completion times over 13 procedures. Figure E.2 illustrates how $CPD(D_{\text{reference}})$ and $CPD(D_{\text{measured}})$ are evaluated. Figure E.3 shows the Monte Carlo approach for showing significance of procedural specific variability. The $CPD(D_{RS \_ref})$ represents the distribution of $D$ values we would expect for showing the extent of inter-procedural variability. If $D_{\text{ref \_meas}}$ is within the 90 percentile limits of $CDF(D_{RS \_ref})$ then there is no evidence of procedural specific variability. This analysis also shows evidence our bootstrapping approach is appropriate as we see correspondence between the analytical $p$-value distribution and the $p$-value distribution evaluated using the Monte Carlo approach. The analytical $p$-value is evaluated directly from the K-S statistic and is a function of $D$ and $N_e$. Figure E.4 shows how $CPD(D_{\text{ref \_meas}})$ is evaluated using a bootstrapping approach and a comparison of the $CPD(D_{\text{ref \_meas}})$ to $CPD(D_{RS \_ref})$ for multiple procedures.
Figure E.2a – The calculation methods used to find \( \text{CPD}(D_{\text{reference}}) \) and \( \text{CPD}(D_{\text{measured}}) \). \( \text{CPD}_j \) and \( \text{CPD}_T \) represent distribution of completion times for a single procedure and all procedures respectively.

Figure E.2b – Cumulative Probability Distribution’s (CPD) of \( D_{\text{reference}} \) and \( D_{\text{measured}} \) for increasing number of procedures (right). From right to left the distributions represent 1, 2, and 3 groupings of procedures.
Randomly sample (without replacement) in lengths $n$

$n = \text{number of procedures}$

Find $D$

Figure E.3a – The calculation method used to find $\text{CPD(D}_\text{RS_ref})$

Figure E.3b – Monte Carlo simulation results from $\text{CPD(D}_\text{reference})$. From right to left the distributions represent 1, 2, and 3 groupings of procedures.
Figure E.4 – The calculation method used to find CPD(D$_{\text{ref meas}}$) (left). The uncorrected $D$ values for $D_{\text{ref meas}}$ and $D_{\text{RS ref}}$ (right).

E.3 Sequencing Comparative Analysis using the $\delta$ Measure

For the purpose of evaluating the difference between event sequences we compared a single event sequence from one procedure ($ES_i$) to a global reference sequence composed from all other procedures ($CES_{\text{all}}$). To distinguish discrepancy between the two event sequences we first calculate the transition probabilities of all possible states as seen from both event sequences and then use a sequence difference measure $\delta$ defined as:

$$\delta = \frac{\sum_{i=1}^{n} (P_{\text{ref},i} - P_{\text{sample},i})^2 w_i}{2 \sum_{i=1}^{n} w_i}$$

(E.3)

where $P_{\text{ref},i}$ and $P_{\text{sample},i}$ are the $i$th transition probabilities of $TPM_{\text{ref}}$ and $TPM_{\text{sample}}$ respectively, $w_i$ represents the frequency of the $i$th event in the reference sequence and $n$ is the total number of transition probabilities.

The sequence difference measure $\delta$ is analogous to our $KS (D)$ statistic as it is non-dimensional and varies on a scale of zero to one (similar to dissimilar). In addition, our measure holds a metric property (i.e. for sequences $S$, $T$, and $U$: $\delta(S,T)>0$, non-negative; $\delta(S,T)=0 \iff S=T$, coincidence; $\delta(S,T)=\delta(T,S)$, symmetry; $\delta(S,T)+\delta(T,U)\leq\delta(S,U)$, triangle (Mannila, 1998)).
E.3.1 Establishing Confidence Bounds

As with the K-S difference measure the sequencing difference measure is based on two finite sequences resulting in an estimate of the difference between the two. We used a bootstrapping approach to assign a confidence interval to our as outlined in Figure E.5.

![Diagram](image)

Figure E.5 – Comparison of a single event sequence to a reference event sequence to give a similarity measure $\delta$. The reference event sequence is based on a collection of event sequences. NOTE: Random Starting Sequence – sequence based on the probability of the first event occurring. Random Transition Sequence – sequence based on weighted transition probabilities of the reference model.

E.3.2 Correction of $\delta$ for Sample Size

As of yet we have not developed a correction for the $\delta$. Preliminary discussions suggest using a Monte Carlo simulation based on artificial data as in the $D_i$ correction.

E.3.3 Reliability / Patient Variability

In this section we develop a method for quantifying the effects of patient variability on surgical event sequencing. The method is analogous to the K-S method developed for evaluating time, kinematic, and joint angle data. The example provided is of tool sequencing for the entire procedure over 15 procedures. Figure E.6 illustrates how $CPD(\delta_{\text{reference}})$ and $CPD(\delta_{\text{measured}})$ are evaluated. Figure E.7 shows the Monte Carlo approach for showing significance of procedural specific variability. Figure E.8 shows how $CPD(D_{\text{ref meas}})$ is evaluated using a bootstrapping approach and a comparison of the $CPD(D_{\text{ref meas}})$ to $CPD(D_{RS ref})$ for multiple procedures.
Figure E.6a – The calculation methods used to find $\delta_{\text{measured}}$ (left) and $\delta_{\text{reference}}$ (right). NOTE: Random Starting Sequence – sequence based on the probability of the first event occurring. Random Transition Sequence – sequence based on weighted transition probabilities of the reference model.

Figure E.6b – Cumulative probability distribution's (CPD) of $\delta_{\text{reference}}$ and $\delta_{\text{measured}}$ for increasing number of procedures (right). From right to left the distributions represent 1, 2, and 3 groupings of procedures.
Randomly sample (without replacement) in lengths \( n \) \( n = \) number of procedures

\[ r \text{ times } r >> 0 \]

Find \( D \)

\[ \text{CPD}(D_{RS_{ref}}) \]

Figure E.7a – The calculation method used to find \( \text{CPD}(D_{RS_{ref}}) \)

Figure E.7b – Monte Carlo simulation results from \( \text{CPD}(\delta_{\text{reference}}) \). From right to left the distributions represent 1, 2, and 3 groupings of procedures.
**Figure E.8** – The calculation method used to find CPD(D<sub>ref meas</sub>) (left). The uncorrected D values for D<sub>ref meas</sub> and D<sub>RS ref</sub> (right).
Appendix F
Collection of Performance Measure Analysis Results

This appendix presents a collection of performance measures analyses evaluated at each level of the hierarchical decomposition. The purpose is to develop a database of performance measures useful for future analysis and reference.

F.1 Phase Level

F.1.1 Time

![Bar chart showing phase level completion times. Categories and typical task times from Traverso et al. (1997).]

**Figure F.1** – Phase level completion times. Categories and typical task times from Traverso et al. (1997).

![Bar chart showing percentage contribution of phases in the total procedure.]

**Figure F.2** – Percentage contribution of phases in the total procedure.
F.1.2 Kinematics

There are no kinematic performance measures at the phase level of the hierarchical decomposition.

F.1.3 Joint Angles

<table>
<thead>
<tr>
<th>Joint</th>
<th>Non-Shifted Data</th>
<th>Shifted Data</th>
<th>Mean Non-Shifted</th>
<th>Mean Shifted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shoulder</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elbow FE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrist FE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrist RU</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forearm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Diagram of joint angles](image)

**Figure F.3** – Ranges of difference measures between joint angle distributions from each individual procedure and the concatenated data from the remaining six procedures, assessed by the Kolmogorov-Smirnov statistic. The confidence intervals marked indicate 90% confidence bound and are based on a bootstrap estimate. Shifted data represents joint angle distributions normalized to have a common mean.
F.1.4 Event Sequencing

Figure F.4 – Tool exchange state diagram. Confidence intervals found using equation D.1.
Figure F.5 – Ranges of tool exchange (procedural and phase level) sequence similarity measures from each individual procedure and the concatenated data from the remaining 14 procedures, assessed by our event sequencing similarity measure. The confidence intervals marked indicate one standard deviation and are based on a bootstrap estimate.

F.2 Stage Level

F.2.1 Time

Figure F.6 – Stage level completion times during the CDD phase.
F.2.2 Kinematics

There are no kinematic performance measures at the phase level of the hierarchical decomposition.

F.2.3 Joint Angles

There are no joint angle performance measures at the stage level of the hierarchical decomposition.

F.2.4 Event Sequencing

Figure F.7 – Stage level state diagram for the CDD phase. Confidence intervals found using equation D.1.
Figure F.8 – Ranges of stage level sequencing (CDD: ie. exploration, dissection, vessel separation, and control bleeding) similarity measures from each individual procedure and the concatenated data from the remaining 14 procedures, assessed by our event sequencing similarity measure. The confidence intervals marked indicate a 90% confidence bound and are based on a bootstrap estimate.

F.3 Task Level

F.3.1 Time

Figure F.9 – Clipping task completion times in different contexts. The confidence intervals marked indicate one standard deviation.
Figure F.10 – Ranges of difference measures between clipping (task and subtask) completion times distributions in different contexts, assessed by the Kolmogorov-Smirnov statistic. The confidence intervals marked indicate a 90% confidence bound and are based on a bootstrap estimate. (Time_context_comp.xls)

F.3.2 Kinematics

Figure F.11 – Ranges of difference measures between kinematic performance measure distributions in different contexts, assessed by the Kolmogorov-Smirnov statistic. The confidence intervals marked indicate a 90% confidence bound and are based on a bootstrap estimate. (Performance_Measures.xls)
Figure F.12 – Range of difference measures between average action membership value distributions in different contexts, assessed by the Kolmogorov-Smirnov statistic. The confidence intervals marked indicate a 90% confidence bound and are based on a bootstrap estimate. (Performance_Measures_Actions1.xls)
F.3.3 Joint Angles

Figure F.13 – Range of difference measures between mean joint angle distributions in different contexts, assessed by the Kolmogorov-Smirnov statistic. The confidence intervals marked indicate a 90% confidence bound and are based on a bootstrap estimate.

Figure F.14 – Range of difference measures between maximum joint angle distributions in different contexts, assessed by the Kolmogorov-Smirnov statistic. The confidence intervals marked indicate a 90% confidence bound and are based on a bootstrap estimate.
F.3.4 Event Sequencing

Figure F.15 – Task level state diagram for cystic duct (left) and cystic artery (right) vessel separation. Confidence intervals found using equation D.1.

Figure F.16 – Ranges of task sequence (vessel separation: ie. clip, clip, clip, cut) similarity measures from each individual procedure and the concatenated data from the remaining 14 procedures, assessed by our event sequencing similarity measure. The confidence intervals marked indicate one standard deviation and are based on a bootstrap estimate.
F.4 Subtask Level

F.4.1 Time

![Percentage Contribution of Subtasks in Clipping Application Tasks]

Figure F.17 – Percentage contribution of subtasks in a clipping application tasks.

F.4.2 – Kinematics

In the interest of conserving space the difference measures for kinematic and average membership value performance measures for the subtask levels have not been provided. The performance measure results at the subtask level are consistent with results at the task level. Subtask difference measure plot can be found in: performance_measures.xls.

F.4.3 – Joint Angles

In the interest of conserving space the difference measures for joint angle performance measures for the subtask levels have not been provided. The performance measure results at the subtask level are consistent with results at the task level. Subtask difference measure plot can be found in: performance_measures.xls.
F.4.4 – Event Sequencing

Figure F.18 – Task level state diagram for subtasks in clipping applications. Confidence intervals found using equation D.1.
Figure F.19 – Ranges of subtask sequence (clipping: ie. approach, tissue manipulation, apply clip, withdrawal) similarity measures from each individual procedure and the concatenated data from the remaining 14 procedures, assessed by our event sequencing similarity measure. The confidence intervals marked indicate one standard deviation and are based on a bootstrap estimate.
Appendix G
BioRome Conference Submission

This document was submitted to and presented at the Fifth International Symposium on
Computer Methods in Biomechanics and Biomedical Engineering in Rome, Italy in

Repeatability of Ergonomic Stress Scores
in Laparoscopic Surgery

Paul B. McBeth¹, Antony J. Hodgson¹, PhD, Alex G. Nagy², MD.

Departments of ¹Mechanical Engineering and ²Surgery,
University of British Columbia, Vancouver, BC, Canada V6T 1Z4

Abstract

Minimally invasive techniques have been rapidly adopted into mainstream surgery over
the last decade, yet the effects of the instruments on surgeon performance, comfort and
safety have not been well evaluated. Ergonomic assessments in minimally invasive
surgery are comparatively rare, despite widespread acknowledgement that strain injuries
do occur to surgeons. An optoelectronic and video motion analysis system is used to
acquire postural measurements at frequencies of ~20 Hz. These measurements are
converted to ergonomic stress scores using a modified RULA method. One expert
surgeon was evaluated in seven clinical laparoscopic cholecystectomies over a period of
four months. We successfully recorded postures at least once per second during 92±6%
of the time the surgeon was performing tissue manipulation tasks. We found the
ergonomic stress scores were comparatively high throughout the procedures, particularly
for the wrist and elbow. The Kolmogorov-Smirnov statistic was used to compare
similarities of joint angle distributions over multiple trials. The results show marked
similarities between joint angle distributions over multiple trials, which suggests
repeatability in our ergonomic assessment method. This system will be a critical
component in validating surgical simulations for use in training and certifying surgeons
and in designing and evaluating equipment.
1. Introduction

Minimally invasive techniques have been rapidly adopted into mainstream surgery over the last decade, yet the effects of the instruments on surgeon performance, comfort and safety have not been well evaluated. Ergonomic assessments in minimally invasive surgery are rare, despite widespread acknowledgement that strain injuries do occur to surgeons. Radermacher et al. (1996) was one of the first to assess minimally invasive surgeries and outlined improvements for workplace design. Berguer showed the postures in laparoscopic surgery are more stressful than in conventional open surgery (Berguer 1997). In this paper we investigate the biomechanical stresses of an expert surgeon over multiple procedures. The objective is to provide an assessment of biomechanical stresses experienced by an expert surgeon over multiple laparoscopic cholecystectomy procedures and to establish the foundation of a method to quantitatively benchmark new tool designs.

2. Methods

2.1 Equipment

An optoelectronic motion analysis system was used in conjunction with video recordings to acquire postural data and tool tip trajectories at frequencies of ~20 Hz during seven live laparoscopic cholecystectomy procedures. For this study, we used a Northern Digital Polaris Hybrid Optical Tracking System capable of tracking the 3D position of both active infra-red light emitting diodes (IREDs) and passive retro-reflective markers with an accuracy of ~0.2-0.3 mm. The Polaris system was connected to a standard PC (800 MHz AMD Duron) running custom-designed data acquisition software written in Matlab. In addition to the optoelectronic system, we recorded video images of the surgery using both a laparoscope and an external camera focused on the surgeon. The images from these two sources were time stamped and recorded onto standard VHS tape; from these images, we could later determine the stage of the operation and correlate it with the detailed motion measurements. All equipment used in the operating room was tested and approved by Vancouver Hospital’s Biomedical Engineering Department and sterilized with ethylene oxide when appropriate.

2.2 Experimental Protocol

One expert surgeon was evaluated in seven clinical laparoscopic cholecystectomies over a period of four months at Vancouver Hospital. The surgical instrumentation used for the study consisted of a standard set of reusable laparoscopic tools designed for cholecystectomy procedures. We attached sterilized marker arrays to the surgeon’s torso and dominant arm (on the proximal and distal forearm and on the hand), as shown in Figure 1(M,R). This set of marker arrays enabled us to track and record the surgeon’s joint angles at the shoulder, elbow, and wrist. Arrays were secured to the surgeon using elastic and Velcro® harnesses. One researcher scrubbed into the surgery and attached the sterilized marker arrays to the surgeon’s hands, lower arms and torso and laparoscopic tools. This researcher remained to assist the surgeon with any adjustments to the marker arrays or support cuffs. Immediately postoperatively, the surgeon’s joint centre locations were identified using a set of standard isolated joint motions. Neutral joint positions were also determined.
2.3 Kinematic Model and Calibration

We used a simplified seven degree of freedom rigid-body model of the upper limb as shown in Figure 1(L). The model allows us to represent shoulder abduction/adduction, flexion/extension, and internal/external rotation, elbow flexion/extension, pronation/supination and wrist flexion/extension and radial/ulnar deviation. Using this model, we applied a kinematic calibration method based on circle and sphere fitting techniques to locate the approximate centre of each of the shoulder, elbow, and wrist joints.

![Figure 1 - Kinematic model of the surgeon's arm (left). Marker array mounting schematic (middle) and as mounted in the operating room (right).](image)

2.4 Stress Scoring System

To rate the stress experienced by a surgeon we used a modified version of the Rapid Upper Limb Assessment (RULA) technique (McAtamney 1993). The RULA technique is a posture evaluation method designed for assessing upper limb postures during light manipulation tasks. Stress exposure is evaluated using body posture diagrams and scoring tables which specify different posture zones. We used a modified version of the RULA scoring system which incorporated continuous scoring measures to improve the sensitivity of our assessments (see Figure 2). Since the standard RULA scales have different ranges for each joint, we normalized all scores such that 0.0 corresponds to the minimum RULA value for a joint and 1.0 corresponds to the maximum RULA value.

![Figure 2 - Modified continuous RULA scoring system and the discrete RULA scoring system for shoulder elevation and wrist flexion/extension.](image)
2.5 Comparative Analysis (Kolmogorov-Smirnov Statistic)

We would expect to see some difference in joint angle distributions due to patient and measurement variability; however, since the surgeon is fully trained we expect that his approach to each patient will be as consistent as possible and we would hope that there is sufficiently little variation between procedures that measurements obtained during a single procedure will be reliable indicators of the physical stress experienced by the surgeon. To demonstrate the repeatability of stress scores for a surgeon over multiple surgeries, we report the differences between the distributions of joint angles obtained in each procedure taken singly and those obtained from the remaining six surgeries. The Kolmogorov-Smirnov statistic, $d$, provides a measure of difference between two distributions and indicates the maximum vertical difference between pairs of cumulative distribution functions; $d$ ranges from 0 (similar) to 1 (different).

2.6 Post-Operative Calculations

Using the correlated video data, we reviewed the entire procedure and divided the procedure into three phases of interest: cystic duct dissection (CDD), gallbladder dissection (GBD), and gallbladder removal (GBR). For each major phase of the surgery, we computed a normalized weighted postural stress score (NWPSS) by averaging the instantaneous scaled RULA scores over the observation period and we report these values averaged over the seven procedures.

3. Results

3.1 Marker Visibility

A certain amount of position data is lost during the course of the surgery due to marker occlusion or internal localization errors. Using the correlated video images we separated manipulation tasks from non-manipulation tasks and found that the majority of missing data could be attributed to marker occlusions during segments not directly related to performing a surgical task. We applied a generalized cross-validation technique to interpolate missing data segments shorter than 0.5 s in duration (expected RMS error < 0.9 mm). During tissue manipulation tasks, we obtained at least one sample per second an average of 92% (SD: 6%) of the time.

3.2 Comparative Analysis

For each joint, we calculated and plotted the cumulative distribution function (CDF) of joint angles during each procedure (see Figure 3 for representative examples). Due to variability in defining the neutral posture for each joint, we shifted each distribution such that they share a common mean. The differences in the distributions are therefore primarily due to differences in shape rather than mean values.

Figure 4 shows the range of $d$ values for each joint resulting from comparing joint angle distributions from each procedure to the concatenated distribution of the remaining six procedures, along with the average $d$ value and standard deviation for each joint. The most proximal joints exhibit the least variability ($d \sim 0.10-0.13$, on average), while the
wrist joint exhibits the most ($d \sim 0.2$, on average). Since the surgeon is fully trained, these values represent the minimum variations we might expect to find from procedure to procedure when using the same surgical tools.

**Figure 3** – Cumulative distribution functions for the elbow flexion/extension and wrist radial/ulnar deviations for all seven procedures.

**KS Statistic for Joint Angle Distributions**

<table>
<thead>
<tr>
<th>Joint</th>
<th>KS Statistic</th>
<th>Mean KS Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shoulder</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elbow FE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrist FE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrist RU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forearm PS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4** – Ranges of difference measures between joint angle distributions from each individual procedure and the concatenated data from the remaining six procedures, assessed by the Kolmogorov-Smirnov statistic. The confidence intervals marked indicate one standard deviation and are based on a bootstrap estimate.

### 3.3 Posture Scores

Using the three operative phases CDD, GBD, and GBR, we computed the modified normalized continuous RULA posture scores for each joint; these are shown in Figure 5. The elbow and wrist joints show significant stress throughout all phases of the procedure, as does the forearm during gall bladder dissection.
4. Discussion

The primary purpose of this study was to assess the stresses experienced by an expert surgeon over multiple laparoscopic cholecystectomy procedures and to demonstrate a new method to assess the potential benefit of new tool designs. By comparing joint angle distributions, we investigated the effect of patient-specific factors on measurements of physical stress. The low values and variance of $d$ for the shoulder, elbow, and forearm suggest that these joint angle distributions are only slightly affected by patient variability. The higher variance of $d$ for wrist flexion/extension and radial/ulnar deviation suggest moderate inter-patient effects.

The ranges of $d$ established in this study will prove useful when evaluating the effect of new tools with improved ergonomics. Such tools, if effective, would produce joint angle distributions characterized by $d$ values typically larger than 0.2 when compared with the distributions found in this study. Similarly, such improvements ought to be reflected in decreased stress scores. Unfortunately, current tools seem to produce significant stress at virtually all joints (excluding the shoulder) throughout all phases of the procedure; this finding is consistent with other reports (e.g., Berguer 1998).

5. Conclusions

We successfully used an automated high-frequency postural measurement system to assess the stresses experienced by an expert surgeon over seven laparoscopic cholecystectomy procedures. Joint stress levels are high in all phases of surgery and the elbow and wrist are significantly more stressed than the shoulder and forearm. Using the Kolmogorov-Smirnov statistic, we created a difference scale to assess the effects of inter-patient variability. This new scale can be used to evaluate new tool designs, validate surgical simulations and assess the performance of surgeons.
6. References

Quantitative Methodology of Evaluating Surgeon Performance in Laparoscopic Surgery

Paul B. McBeth¹, Antony J. Hodgson¹, PhD, Alex G. Nagy², MD, Karim Qayumi², MD PhD

Departments of ¹Mechanical Engineering and ²Surgery, University of British Columbia, Vancouver, BC, Canada V6T 1Z4

Abstract

Quantitative performance and skill assessments are critical for evaluating the progress of surgical residents and the efficacy of different training programs. Current evaluation methods are subjective and potentially unreliable, so there is a need for objective methods to evaluate surgical performance. We identify a feasible method to measure kinematic data in the live operating room setting and to assess the repeatability of an analysis method based on a hierarchical decomposition of surgical tasks. We used an optoelectronic motion analysis system to acquire postural data and tool tip trajectories of one expert surgeon over a period of four months. To assess repeatability of performance measures, we created a hierarchical decomposition diagram describing the procedure in terms of surgical tasks, tool sequences and fundamental tool actions. From the kinematic data, we extracted characteristic measures of individual tool actions and compared these measured distributions using the Kolmogorov-Smirnov statistic. The comparisons of distributions show consistent performance over time by a trained surgeon and little effect from patient variability, and so are likely reliable measures of performance. An expanded set of reliable kinematic measures will form the basis for quantifying surgical skill and should be useful in validating surgical simulations for use in training, certifying surgeons and designing and evaluating new surgical tools.
1. Introduction

Although minimally invasive techniques have been rapidly adopted into mainstream surgery over the last decade, developments in teaching and evaluation methods have not kept pace. Quantitative performance and skill assessments are critical for evaluating the progress of surgical residents and the efficacy of different training programs. They are also important in assessing new surgical tool sets and techniques. Current evaluation methods are subjective and potentially unreliable (Rosser, 1998) so there is a need for objective methods to evaluate surgical performance. One valuable approach, which was recently demonstrated on a porcine model, is based on an analysis of tool tip force/torque signatures (Rosen, 2001). In this paper, we discuss a complementary approach which incorporates kinematic features of both tool motion and surgeon posture. The purposes of this study were to identify a feasible method to measure kinematic data in the live operating room setting and to assess the repeatability of an analysis method based on a hierarchical decomposition of surgical tasks. An abbreviated sample of the kinematic data results is presented due to limited space.

2. Methods

2.1 Equipment

An optoelectronic motion analysis system and video recordings were used to acquire postural data and tool tip trajectories at frequencies of ~20 Hz. For this study, we used a Northern Digital Polaris Hybrid Optical Tracking System capable of tracking the 3D position of both active infra-red light emitting diodes (IREDs) and passive retro-reflective markers with an accuracy of ~0.2-0.3 mm. The Polaris system was connected to a standard PC (800 MHz AMD Duron) running custom-designed data acquisition software written in Matlab.

In addition to the optoelectronic system, we recorded video images of the surgery using both a laparoscope and an external camera focused on the surgeon. The images from these two sources were time stamped and recorded onto standard VHS tape; from these images, we could later determine the stage of the operation and correlate it with the detailed motion measurements. All equipment used in the operating room was tested and approved by Vancouver Hospital’s Biomedical Engineering Dept. and sterilized with ethylene oxide when appropriate.

2.2 Experimental Protocol

One expert surgeon was evaluated in seven clinical laparoscopic cholecystectomies over a period of four months at Vancouver Hospital. We attached sterilized marker arrays to the surgeon’s torso and dominant arm (on the proximal and distal forearm and on the hand), as shown in Figure 1(L,C). This set of marker arrays enabled us to track and record the surgeon’s joint angles at the shoulder, elbow, and wrist. Arrays were secured to the surgeon using elastic and Velcro® harnesses. Custom-designed Multi-Directional Marker Arrays (MDMAArray) were attached to the each laparoscopic tool handle used by the dominant hand, as shown in Figure 1(R). A planar marker array was attached to the
non-dominant hand tool. The MDMArrays were equipped with a quick release clip to allow easy attachment and removal. The Polaris system is capable of tracking six different laparoscopic tools.

One researcher scrubbed into the surgery and attached the sterilized marker arrays to the surgeon’s hands, lower arms and torso and laparoscopic tools. This researcher remained to assist the surgeon with any adjustments to the marker arrays or support cuffs. Immediately postoperatively, the surgeon’s joint centre locations were calibrated using a set of standard isolated joint motions. The tip position of each surgical tool was found using a sphere-fitting calibration procedure in which the tip was placed in a small depression on a fixed surface while the tool was manipulated about that point.

2.3 Missing Data

Optoelectronic systems have inherent line-of-sight limitations causing segments of missing data. Custom-designed tool marker arrays with multiple faces visible from several directions (the MDMArrays mentioned above) were developed to minimize these problems. To establish the feasibility of using an optoelectronic system for tool tip and posture tracking in the operating room, we measured the frequency of marker occlusions and estimated the errors associated with interpolating missing data.

Figure 1 - Model of the surgeon’s arm (L). Marker array mounting (schematic and as mounted in the operating room (C)). Multi-Directional-Marker-Array (MDMA) (schematic and as mounted on a laparoscopic surgical tool (R). The values labelled on each marker represent the percentage of time it was visible during manipulation segments (%).
2.4 Hierarchical Decomposition

Many performance measures (e.g., distance for a tool tip relocation maneuver) can only be defined for low-level movement segments. To assess repeatability of such performance measures, we therefore created a hierarchical decomposition describing the procedure in terms of surgical phases and stages, tool tasks and subtasks, and fundamental tool actions (see Figure 2). This five-level hierarchical decomposition provides the foundation for a quantitative analysis of surgeon performance since it allows us to identify related tool actions performed at different points during the procedure and consolidate the associated performance measures for analysis.

**Figure 2 - Laparoscopic Cholecystectomy Hierarchical Decomposition.**

2.5 Performance Measures

From the kinematic data associated with each subtask, we extracted characteristic measures of individual tool movements; for example: duration, mean and peak tool tip velocity, peak acceleration, jerk cost, straight-line deviations. In additional to the kinematic performance measures we also investigated mean membership values throughout a subtask in each of 12 prototypical actions (e.g., reach, sweep, idle, etc.); membership is expressed using fuzzy membership functions based on instantaneous tool tip position. We selected an example task (applying a clip) which is performed comparatively frequently (at least 6 times per procedure) and compared the characteristic measures associated with its component actions across the 35 clipping segments recorded.
2.6 Comparative Analysis (KS Statistic)

To demonstrate the reliability of the proposed analysis procedure, we must show the performance measures are comparatively unaffected by variations between patients yet sensitive to real differences such as training effects. We assess similarity by comparing distributions of performance measures obtained from multiple executions of particular subtasks in two different contexts (e.g., comparing clip application times from the first three procedures to those from the last three could reveal a learning effect). To compare distributions, we use the Kolmogorov-Smirnov (KS) statistic \(d\), which represents the maximum vertical difference between cumulative distribution functions (see Figure 3); this measure ranges from 0 (similar) to 1 (different). The associated \(p\) value expresses the probability that the two measured distributions arise from the same underlying distribution.

We performed several comparative analyses related to the clip application subtask. Most commonly, clips are used in the vessel separation (CD and CA) stage of the procedure; however, they are also occasionally used to control bleeding throughout the procedure. We would expect that there would be no significant difference between performance measures taken from alternate clip applications, as both patient variability and any learning effects will be equally represented in the two data sets. Since the surgeon is fully trained, we would hope for little variation between the first four and the last three procedures studied; any differences will be primarily due to patient variability and the results will represent the reliability of our analysis procedure. Finally, we prospectively investigated several other divisions of the data set in hopes of demonstrating the existence of significant and interesting differences; for example, we considered whether there might be differences between clips applied for: vessel separation vs. control of bleeding; separating the cystic duct vs. the cystic artery; first placement vs. subsequent placement on the same vessel.
3. Results

3.1 Missing Data

A certain amount of position data is lost during the course of the surgery due to marker occlusion or internal localization errors. Figure 1(L) illustrates the marker array visibility averaged over the seven procedures multiple trials for individual markers. Using video analysis to separate manipulation from non-manipulation tasks, we found that the majority of missing data could be attributed to marker occlusions during segments not directly related to performing a surgical task (e.g. changing instruments). During tissue manipulation tasks, we had complete joint angle data 80% (SD: 10%) of the time and at least one sample per second 92% (SD: 4%) of the time. The trackability of the dominant hand tool depends on which tool is being used (we tracked the tool 78% (SD: 12%) of the time during the clipping task; 12 of the 45 tasks had no lost data and 20 successfully recorded more than 90% of the samples).

We used a generalized cross validation (GCV) filtering technique (Woltring, 1986) to find an optimal smoothing parameter for fitting a quintic spline to position data. The data was resampled at constant intervals and missing data segments interpolated. The error associated with interpolating across varying gap sizes was calculated and shown in Figure 4. An RMS error of 1mm was chosen as an acceptable error which suggests that a maximum gap size of 0.5 sec (10 samples) can be interpolated. We were able to reduce the amount of missing joint angle data by 20% (SD: 2%) using our interpolation technique with a maximum 1mm RMS error. Using the same interpolation, we reduced the amount of missing data for the right and left hand tool data by 6% (SD: 2%) and 14% (SD: 3%) respectively.

![Graphical representation of missing data interpolation](image)

**Figure 4** - Graphical representation of missing data interpolation (L). RMS error associated with missing data gap size (R).
3.2 Hierarchical Decomposition

The hierarchical decomposition of the procedure provides an organizational framework for the kinematic and event sequencing analysis of different surgical tasks. The decomposition design was based on the standard elements of laparoscopic cholecystectomies and was developed in consultation with two expert surgeons; the decomposition approach can be readily adapted to other laparoscopic procedures. The decomposition has five levels, each containing more specific details than the previous. The phase level (1) outlines the global goals of the procedure. The stage level (2) outlines local goals required to complete each phase. The task level (3) involves the use of a single tool. The subtask level (4) describes how the surgical tool is moving inside the patient. Finally, actions (5) describe kinematic features of stereotyped fundamental movements such as reaching or sweeping. Figure 2 shows the various levels of the hierarchical framework.

3.3 Performance Measures

Performance measures and mean action membership values were evaluated at the task and subtasks levels for the clip application task. Figure 5 illustrates a limited collection of performance measures and their respective difference scores for clipping tasks.

Figure 5 – Kolmogorov-Smirnov difference statistic of performance measures and mean membership action variables.
4. Discussion

We successfully recorded tool tip position 82% (SD: 22%) of the time the surgeon was performing tissue manipulation tasks and posture data at least once per second 92% (SD: 4) of the time. Using a generalized cross validation interpolation technique, we were able to interpolate one second gaps with estimated RMS errors less than 0.64 mm. Using the KS statistic the small difference between the time measures in the odd and even trials and 1st half / 2nd half trials suggests there are no obvious learning effects and no significant effect of patient variability. This is confirmed by the kinematic measures, although the difference measures are higher in large part because of the limited sample size. In addition, there appears to a significant difference between clips applied to control bleeding and clips applied for vessel separation.

5. Conclusion

We used an automated high-frequency tool tracking and postural measurement system to identify kinematic features of specific surgical actions of laparoscopic cholecystectomies as defined in a hierarchical decomposition diagram. The results demonstrate that an optoelectronic and video motion analysis, particularly in conjunction with a missing data interpolation technique, is a reasonable method for recording kinematic data in the operating room, despite occasional periods of missing data. Additional data is required to decrease the confidence intervals of the reported difference measures. The comparisons of distributions of those performance characteristics of surgical actions which we evaluated show consistent performance over time by a trained surgeon, and so are likely reliable measures of performance. An expanded set of reliable kinematic measures and data will form the basis for quantifying surgical skill and should be useful in validating surgical simulations for use in training, certifying surgeons and designing and evaluating new surgical tools.

6. Acknowledgements

We thank the operating room staff at Vancouver Hospital and Health Sciences Centre for their cooperation and assistance. We also thank the Natural Sciences and Engineering Research Council of Canada (NSERC).

7. References

Appendix I S.A.G.E.S. Conference Submission

This abstract has been accepted for poster presentation at the Society of American Gastrointestinal Endoscopic Surgeons (SAGES) 8th World congress of Endoscopic Surgery in New York, USA in March, 2002.

Intra-surgeon Variability in Motor Task Performance in Laparoscopic Surgery

Paul B. McBeth¹, Antony J. Hodgson¹, PhD, Alex G. Nagy², MD, Karim Qayumi², MD PhD

Departments of¹ Mechanical Engineering and ²Surgery, University of British Columbia, Vancouver, BC, Canada V6T 1Z4

Abstract

Background: Current methods of evaluating the skills of surgical residents are subjective and potentially unreliable, so there is a need for objective methods to monitor their training. The purpose of this study is to test the intra-surgeon reliability of a proposed quantitative skill assessment method based on surgical tool kinematics.

Methods: One expert surgeon performed seven clinical laparoscopic cholecystectomies over a period of four months. Using an optoelectronic motion analysis system we acquired tool tip trajectories at frequencies of ~20 Hz. A hierarchical decomposition matrix was used to segment the procedure into specific surgical actions and extracted characteristic measures of these individual actions (e.g., duration, mean tool tip velocity, etc.). We selected an example sequence (applying a clip) and compared the characteristic measures associated with its component actions across the seven procedures recorded.

Results: We found no statistical similarity difference between the completion time distributions in any of the four contexts examined: (1) first 4 trials vs. last 3 trials: the probability of these segments arising from the same distribution is p=0.30, and a percent difference in mean completion time, which would have been deemed significant of (d)=5.4%, (2) cystic artery vs. cystic duct: p=0.1, d=22.3%, (3) 1 clip vs. subsequent clips (cystic artery): p=0.94, d=9.3%, and (4) 1 clip vs. subsequent clips (cystic duct): p=0.67, d=20.8%. Results for trajectory length and tip velocity are similar.

Conclusions: We used an automated tool tracking system to identify kinematic features of specific surgical actions during laparoscopic cholecystectomies. The results suggest an expert surgeon performs specific tasks quite consistently from case to case and the performance measures are only modestly sensitive to specific patients.