Quantifying Surgical Skill

By

David Paul Azari

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Industrial and Systems Engineering)

At the

UNIVERSITY OF WISCONSIN-MADISON

2018

Date of final oral examination: 08/15/2018

The dissertation is submitted to the following members of the Final Oral Committee:

Robert G. Radwin, Chair, Professor, Industrial and Systems Engineering
Caprice C. Greenberg, Professor, Industrial and Systems Engineering, Surgery
Yu H. Hu, Professor, Electrical and Computer Engineering
Carla M. Pugh, Professor, Industrial and Systems Engineering, Surgery
Douglas A. Wiegmann, Assistant Professor, Industrial and Systems Engineering

Dedication

This dissertation was made possible by Melissa M. Azari, wife.

Abstract

Surgical performance lacks an objective framework to promote quantifiable skill assessment. Expert surgeons naturally recognize features of good performance when they see it but quantifying and consistently reproducing such features remains an open problem. Statistical modeling through computer vision of hand motion – enabled by increasing and easily scalable access to digital video records – may hold the key to quantify surgical performance without depending on robotic or sensor-based systems. This dissertation investigates how video and computer vision of surgical hand motion can effectively quantify performance in and out of the operating room.

The specific aims of this dissertation use video to: (1) identify kinematic features of hand motion associated with increasing clinician experience, (2) train machine learning algorithms to identify periods of suturing and tying from a continuous video record, (3) automatically predict expert-rated performance of common benchtop suturing tasks, and (4) examine the validity of expert-rated performance predictions in live operating room procedures. A new model defining surgical skill terminology is also proposed to ensure consistency describing surgical skill, and to frame future study. Each portion of this work takes a necessary step to enable continuing surgical skill analysis utilizing digital video.

These aims are accomplished from video analysis of 92 surgeons and students performing common suturing and tying tasks. Video (9 hours and 32 minutes) of clinicians of varying experience suturing on foam and pig feet were collected and analyzed. Residents exhibited greater movements with their dominant hands than medical students, while reducing the path length needed to complete the same task. Experience as an attending surgeon was associated with increased or similar cycle frequency, but reduced acceleration and path length per cycle of the non-dominant hand compared to residents. These results suggest that early increases in tenure

are associated with more purposeful dominant hand use, while gains in residency and through attending roles promote simple movements and conserve energy where possible.

In the second section, video records served as fodder to train a series of machine learning algorithms to recognize suturing and tying tasks from a continuous record. A Hidden Markov Model (HMM) predicted 79% of states for a reserve set of participants, and reasonably predicted the completion rate of each participant: slope = 0.88, intercept = 0.03, correlation = 0.83, R^2 = 0.72.

In the third and fourth phases of this work, experts rated performance for 219 clips using a custom program. Four visual-analog rating scales developed in previous work (Azari et al., 2017) were used: fluidity of motion, motion economy, tissue handling, and hand coordination. Motion records of the hands were then used to predict the expert ratings of each scale. Fluidity of motion provided the best prediction for expert-rated scores (slope = 0.71, intercept = 1.98, R^2 = 0.77, correlation = 0.88, R^2_{pred} = 0.70) and extrapolated well to video clips (n = 48) collected in the operating room (slope = 0.83, intercept = 1.75). Motion economy provided a good relationship between predicted and expert rated scores (slope = 0.65, intercept = 2.36, R^2 = 0.66, correlation = 0.81, R^2_{pred} = 0.61) and extrapolated moderately well to the operating room (slope = 0.73, intercept = 2.04). Both models were sensitive (R^2 = 0.55, 0.49) to contextual features of the operating room like changing postures and false starts suturing in friable tissues.

This research provides a timely model of surgical skill terminology, extends automatic segmentation of surgical video, and completes the first empirical study extrapolating automatic, video-based predictions of surgical performance from benchtop settings to the real-world setting of the operating room. These developments could be used to build a video-based formative assessment and feedback tool, aimed at quantifying performance throughout a surgeon's career.

Acknowledgements

Completing this dissertation would not have been possible without the help of many people. I would like to express my gratitude and thanks to my advisor and mentor, Dr. Robert Radwin, for his guidance and support throughout my graduate career. I would also like to thank my committee members, Drs. Caprice Greenberg, Yu Hen Hu, Carla Pugh, and Douglas Wiegmann for their valuable discussions, input and expertise. Thank you also to Brian Le, Brady Miller, Reginald Bruskewitz, Jacob Greenberg, Kristin Long, and Amber Shada for your time and patience in helping to facilitate this work. Thank you to Drs. John Pfotenhauer, Richard Halverson, and Ann O'Rourke for your encouragement and perspective.

Additional thanks to the other members of the Occupational Ergonomics and Biomechanics Lab, particularly Oguz Akkas and Renee Greene, whose daily friendship and kindness I am sure to miss, and to Eric Chen and Karen Chen, whose paths I am honored to follow.

I would also like to express my gratitude to Michael Rios and Dennis Choi – "ye olde housemates" – who lifted my spirits. My in-laws Michael and Diane Warner, Dan and Mary, for being so welcoming. My parents Aaron and Anita Azari, and siblings Abby, Alyssa, Sam and nephew Timothy for their continual support and love. My extended family for a long-lived commitment to education and learning. And to my wife Melissa, the best unintended outcome of attending graduate school.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

ABIM: American Board of Internal Medicine
ACGME: Accreditation Council for Graduate Medical Education
AERA: American Education Research Association
AIC: Akaike Information Criterion
ANOVA: Analysis of Variance
CAE: Computer Assisted Evaluation
CCHs: Chain Code Histograms
CPU: Central Processing Unit
CSMs: Common String Models
DFA: Detrended Fluctuation Analysis
DTW: Dynamic Time Warping
GOALS: Global Objective Assessment of Laparoscopic Skills
GPU: Graphical Processing Unit
GRS: Global Rating Scales
HMM: Hidden Markov Models
HSD: Honestly Significant Difference
ICC: Intra-class Correlation Coefficient
ICSAD: Imperial College Surgical Assessment Device
IOM: Institute of Medicine
k-NN: k-Nearest Neighbors
LASSO: Least Absolute Shrinkage and Selection Operator
MSE: Mean Square Error
MVTA: Multimedia Video Task Analysis
NLP: Natural Language Processing
OR: Operating Room
OSATS: Objective Structured Assessment of Technical Skills
OSANTS: Objective Structured Assessment of Non-Technical Skills
ROI: Region of Interest
ROVIMAS: Robotic Video and Motion Analysis Software
SVM: Support Vector Machines
UPOA: Unstable Periodic Orbit Analysis
WHO: World Health Organization

Introduction

Background

Improved surgical skill decreases the length and variability of operations (Carty, Chan, Huckman, Snow, & Orgill, 2009), frees up cognitive resources (O'Neil, Perez, & Baker, 2014) and promotes better patient outcomes (Birkmeyer et al., 2013). After years of deliberate practice, surgeons hone and test their skills during difficult cases; building "surgical wisdom" (Francis, 2009) and sharing "war stories" (Y. Y. Hu et al., 2012) of what went right (and wrong) along the way. Attendings draw a resident's attention to important cues and provide valuable feedback (Hauge, Wanzek, & Godellas, 2001), while tailoring interaction to guard the patient's safety (Glarner et al., 2017). The lessons and patterns observed over time, commonly called "illness scripts" (Schmidt, Norman, & Boshuizen, 1990), help to construct the expectations of future surgeons and promote readiness to engage in stressful and demanding situations (O'Neil et al., 2014). Despite such rigorous training, however, Mattar et. al., (2013) found lackluster operative autonomy among graduated residents, with many (56%) not able to suture effectively and needing "remedial training" (Bell, 2009). There is also little emphasis on continuing evaluation of surgical skills for attending clinicians, to facilitate professional transitions (Alleman & Al-Assaf, 2005). Objective performance assessment, supported through computer vision of video records, may be able to address these challenges.

Growing ability to integrate surgical information – termed "surgical data science" (Maier-Hein et al., 2017) – and quantify surgical performance with engineering tools in particular (Rutherford, D'Angelo, Law, & Pugh, 2015), are promoting continued development of objective computer-aided technical skill evaluation (OCASE-T) (Vedula, Ishii, & Hager, 2017). The majority of these efforts examining psychomotor skills rely on sensors and markers; recording hand movements, forces, and joint angles or orientations. The publicly accessible

Gesture and Skill Assessment Working Set (JIGSAWS) (Gao et al., 2014) and Robotics Video and Motion Assessment Software (ROVIMAS) (A. Dosis et al., 2003) utilizing the da Vinci robot-assisted platform, are good examples. Hand movements measured through the Imperial College Surgical Assessment Device (ICSAD) also have an impressive track record observing differences between clinician experience (Corvetto et al., 2017; Datta, Mackay, Mandalia, & Darzi, 2001). These advances require access to sensor and data collection systems, limiting their feasibility in open surgical settings.

Our approach, in contrast, uses computer processing of digital video to capture surgeon hand motions. This video is easy to collect, cheap, and scalable even in remote and difficult-to-access areas. Using computer vision to automatically deconstruct surgical video and predict performance, however, remains understudied. In previous work, our group has used video recording in the operating room to: (1) quantify differences in hand motion while attendings and residents conduct the same task (Frasier et al., 2016; Glarner et al., 2014; Radwin et al., 2014), and (2) predict expert ratings of surgical performance during short clips of tying and suturing maneuvers (Azari et al., 2017). These studies, despite their success, were limited by range of clinician experience and setting. They were not able to generalize video-based predictions of performance across a surgical career or extrapolate scores between repeatable benchtop simulations and real-world settings. This dissertation extends the existing body of work to produce novel, computational models of surgical performance across a range of clinician experiences and settings, while improving automatic deconstruction of surgical video into discrete periods of suturing and tying.

Research Questions

This dissertation addresses the following research questions:

1) How does experience impact features of observable hand motion?
2) Can common machine learning techniques classify maneuvers from uninterrupted video with similar accuracy as robotic platforms?
3) How well can features of hand motion predict expert-rated surgical performance during common benchtop suturing tasks?
4) How well do automatic predictions of surgical performance in benchtop settings extrapolate to the operating room?

Dissertation Structure

Chapter 1 develops a nascent model to define common surgical skill terminology to support and frame the remaining chapters. Chapter 1 also discusses existing surgical assessment methods and current frameworks of expertise underpinning surgical skill development. Chapter 2 summarizes observed differences between students and clinicians of varying experience while completing common benchtop suturing tasks on foam, pig feet, and bowel. Chapter 3 explores automatic segmentation of surgical video into discrete periods of suturing and tying using common machine learning techniques in benchtop settings. Chapter 4 uses expert-ratings to automatically predict performance of clinicians of varying experience while suturing on foam. Chapter 5 tests how well prediction models of performance on benchtop simulations extrapolate to open procedures in the operating room. The last portion of this dissertation summarizes future directions of this work, presents several suggestions for optimal video collection and processing in healthcare settings, and proposes a conceptual software interface, which could help students and residents identify their strengths and weaknesses in completing common benchtop suturing tasks.

1. In Search of Characterizing Surgical Skill

1.0 Manuscript Information

This manuscript will be submitted to *The Journal of Surgical Education.*

1.1 Abstract

Objective surgical skill analysis depends on consistent definitions of terms like performance, expertise, experience, aptitude, ability, competency, and proficiency. This paper provides a discussion of surgical skill terminology and proposes a set of unique definitions to facilitate shared understanding among efforts to quantify surgical skill. A new model is proposed to cement a common and consistent lexicon for future skills analysis and describe a surgeon's performance throughout their career.

1.2 Introduction

Inconsistent use of surgical skill terminology is pervasive. Common descriptors like superior performance (or elite performance), aptitude and ability, competency and proficiency, mastery, expertise and experience commonly lack unique interpretation. Experience is sometimes used as a proxy for expertise, but also as a signal of professional status, without supporting evidence of best practices. Laufer and colleagues (2016), for instance, observed that clinicians of similar experience exhibit different approaches (sometimes radically) while completing simulated clinical breast exams. Should physicians with the greatest experience be considered experts even if – as Choudhry and colleagues (2005) found – greater experience doesn't always produce greater quality of care? Or, are only some experts whose patients achieve better outcomes exhibiting what we would think of as "truly expert" behaviors?

Uncovering features to accurately describe surgical expertise is similarly an intricate challenge. Advanced performance in the operating room is known to depend on combining various skill sets (Madani et al., 2017; S Yule, Flin, Paterson-Brown, & Maran, 2006), for which

any objective assessment must be tailor made (Yule & Paterson-Brown, 2018) and rigorously tested (Jelovsek, Kow, & Diwadkar, 2013). Variable terminology hinders this effort, and makes it more difficult to validate assessments in line with robust evidentiary requirements of modern frameworks (Cook & Reed, 2015; Cook, Zendejas, Hamstra, Hatala, & Brydges, 2014). The rise of "surgical data science" (Maier-Hein et al., 2017) and engineering approaches to quantify surgery (Rutherford et al., 2015), in part through simulation (Scott et al., 2008; Vedula et al., 2017), offers the opportunity to address boundary conditions of amorphous terms such as "expert" through quantifying behavior. This paper explores common surgical skills terminology and proposes a lexicon encourage reproducibility among future studies of surgical performance.

1.3 Background

1.3.1 What is surgical skill?

Surgical skills are commonly split into either technical or non-technical categories (Yule et al., 2006), despite known impacts of non-technical skill on technical performance (Hull et al., 2012). It is also widely accepted that "operative skills" are not just technical in nature (Bell, 2009). Still this artificial bifurcation has helped to frame studies examining hand motion (Azari et al., 2015, 2017; Datta, Chang, Mackay, & Darzi, 2002; Frasier et al., 2016; Glarner et al., 2014; Radwin et al., 2014), errors and error management strategies (Law Forsyth et al., 2017; Nathwani et al., 2017; Regenbogen et al., 2007; Rogers et al., 2006), cognitive readiness (O'Neil et al., 2014), decision making (Pugh & DaRosa, 2013) and communication and teamwork (Dedy, Fecso, Szasz, Bonrath, & Grantcharov, 2016; Moorthy, Munz, Adams, Pandey, & Darzi, 2005; Wiegmann, ElBardissi, Dearani, Daly, & Sundt, 2007).

In a discussion of non-technical skills, Yule et al., (2006) proposed the interrelated category of cognitive skills, to better describe features of surgical performance such as mental readiness, decision making and situational awareness. This deconstruction is additionally

supported by surveys of master surgeons (in this case, defined as those with high peer rankings and consistent involvement as trainers), who describe cognitive factors, innate dexterity, and personality as "important attributes" of surgical competence (Cuschieri, Francis, Crosby, & Hanna, 2001). Greenberg et al. (2015) advocates that cognitive skills should also be integrated as part of the Wisconsin Surgical Coaching Framework (Figure 1).

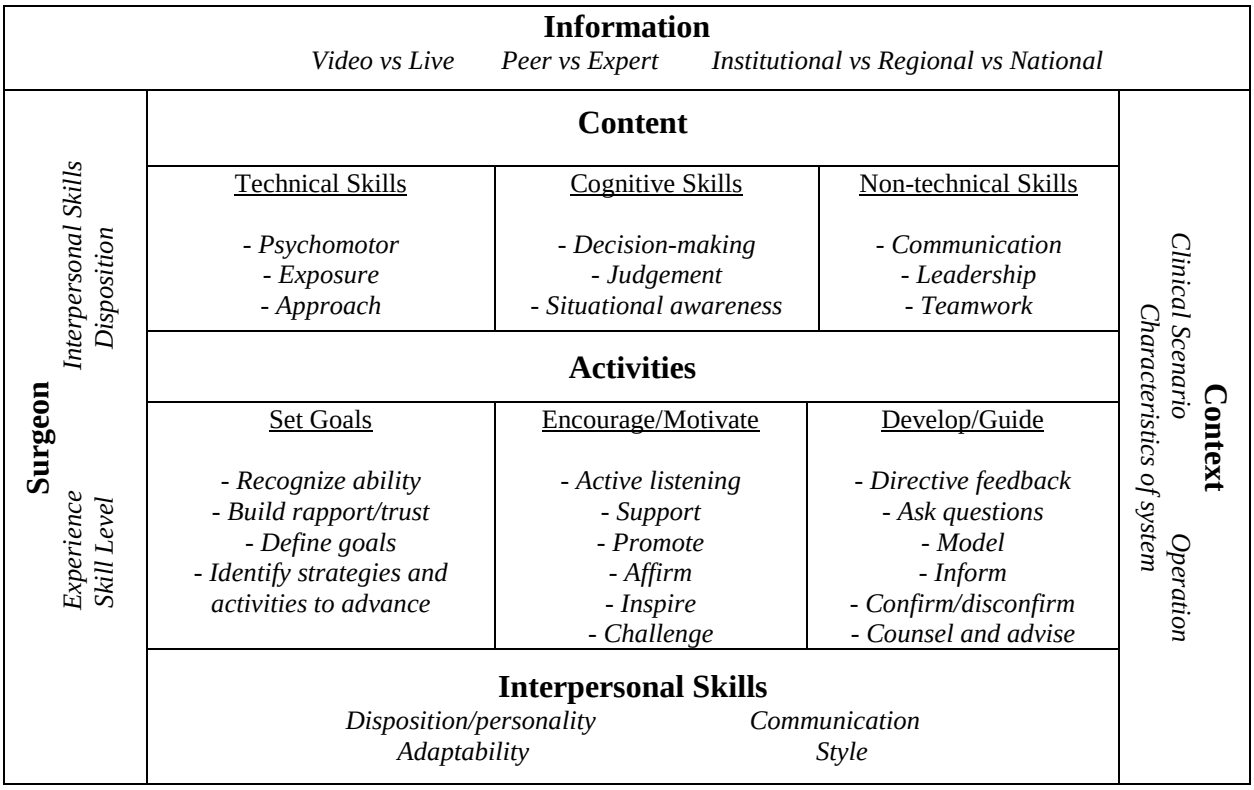


Figure 1: Wisconsin Surgical Coaching Framework (Adapted from Greenberg et al., 2013)

Madani and colleagues (2017) have since developed a novel interoperative performance framework, composed of five inter-related performance domains: psychomotor skills (i.e. technical performance), declarative knowledge (i.e. recitable facts acquired outside the operating room), interpersonal skills (i.e. teamwork, leadership), personal resourcefulness (i.e. self-awareness and metacognition), and advanced cognitive skills (i.e. planning, error recovery) (Madani, as summarized by Perdanasari & Hollier, 2017). This framework provides an excellent description of various domains or "competencies" in which surgeons perform. It facilitates continuing ontological understanding of a skill's "microstructure" – a necessary first step in

designing “deliberate practice activities that allow performers to stretch their performance to a higher level” (Ericsson, 2005, p. 237), and helps to incorporate the “reasoning and motivation” behind successful operations in challenging professional settings (Ginsburg, Regehr, & Lingard, 2004).

Madani and colleagues maintain that “competence has yet to be defined to a level that allows credentialing and licensing bodies to ascertain whether or not an individual has achieved the standards deemed to represent competent performance.” In other words, it remains a challenge to establish “pass-fail standards,” critical parts of competence-based education (Reznick & MacRae, 2006). They are also careful not to refer to their framework as a “skills framework,” presumably to avoid any confusion or contradiction between thinking of knowledge and performance as a skill or vice-versa. Rather, Madani’s work presents a series of performance domains, the creative and fluid synthesis of which, through deliberate practice and rehearsal, may characterize surgical expertise. The authors adroitly navigate the difficulty in defining such terms and focus on the underlying surgical behaviors. They redirect emphasis from examining “skills” to interrelated “performance domains” that are widely applicable across procedures.

1.4 Terminology

Using surgical skills terminology to describe how a surgeon’s performance changes throughout their career is an intricate puzzle. A trainee’s aptitude in one domain, for instance, hinges on a “working definition of superior surgical performance” (Graham & Deary, 1991). Yet what it means to be superior, generally considered a hallmark of medical expertise, presupposes measurement and attainment of competency (Charness & Tuffiash, 2016). Expertise, meanwhile, is interwoven with notions of both performance and skill. Murinson, Agarwal, & Haythornthwaite (2008), for example, frame expertise as the aggregation of “essential skills” (p. 975); while Krawczyk and colleagues (2013),

argue that expertise exists as a set of “exceptional skills” (p. 364) that can be measured and compared in laboratory tasks. In the same vein, Ericsson and Charness (1994) describe experts as natural outliers: “performing at least two standard deviations above the mean level in the population” (p. 731).

The implicit assumption that performance can be sufficiently (1) observed and (2) quantified to sort or rank performance in a meaningful way remains dubious: “[n]o single assessment method can provide all the data required for judgement of anything so complex as the delivery of professional services by a successful physician” (Miller, 1990). Instead, the Accreditation Council for Graduate Medical Education’s (ACGME) Competency Based Medical Education (CBME) milestones approach (2013), continues to promote operative autonomy through guidelines to document and show performance as a precursor. Any valid assessment to demonstrate these skills must be grounded within a robust basis of evidence (Cook, Brydges, Ginsburg, & Hatala, 2015; Cook et al., 2014; Kane, 2006, 2013). Clear and consistent skill terminology will promote better evaluation and eventual application of assessment tools.

1.4.1 Performance

Like other performance domains (music, athletics, for example) surgical performance is repeatedly created anew at each opportunity, where contextual factors can change rapidly (Davids, Button, & Bennett, 2008). Subsequently, and stemming from notions that there exists a “maximal” level, performance is a dynamic, temporary, and alterable characteristic. Deliberate practice improves performance over time (Ericsson, 2004), but performance is also subject to the context of the surgery (Feltovich, Ford, & Hoffman, 1997) and various factors within the work-system (Francis, 2009). Even those most practiced surgeons are not immune from committing errors, or not managing errors

properly. Surgical performance thus represents the observable quality of a sequence of surgical actions at a specific point in time. It is possible (albeit unlikely) for a novice to outperform an expert, or an expert to underperform relative to their position. Such situations would be the exception, rather than the rule.

Defining maximum performance has a long history of debate. Francis Galton – contemporary of Charles Darwin, and who is generally credited with developing both "nature vs nurture" and "eugenics" terms – argued in *Hereditary Genius* (1904) that maximal performance is a rigid and individually determined limit of one's genetic potential. Although he acknowledged that practice improves performance, Galton argued that "genius" depended foremost on your family tree. Later, Snoddy (as cited by Stratton, Liu, Hong, Mayer-Kress, & Newell, 2007) developed the now "ubiquitous" power law of learning, composed of the distinct cognitive, associative and autonomous classical stages of skill acquisition (Anderson, 1982; Fitts & Posner, 1967). Galton's concept of an individual limit was re-envisioned as task-based performance limit – the asymptote of a "learning curve" governed by a power law. In an invited address to *Academic Medicine*, K. Anders Ericsson expands this definition and argues that achieving an "expert" level of performance hinges on intentional and deliberate practice, implemented over long periods (Ericsson, 2004). Davids (2008) reflects Ericsson's productive framework, in describing that "the power law of practice simply states that performance improves with practice, although there are eventual physical limits to this relationship."

Even though these definitions have changed, surgical performance can be thought of as a temporary snapshot into the observable skills a surgeon brings to bear within a given situation. Performance can improve due to amount and style of deliberate practice but remains bounded to some asymptotic limit. Over longer periods, performance can

improve or decline as surgeons age or switch to different types of operations. While transitioning to military service, for example, surgeons struggle to adapt their specialty based expertise to new challenges such as truncal hemorrhage or skeletal reconstruction from penetrating injuries (Kelly et al., 2008; Tyler, Clive, White, Beekley, & Blackbourne, 2010), while their other clinical skills, especially laparoscopy (Perez et al., 2013), decay. Given the intense training required to achieve high performance in any surgical task, there is an ever-present interest in testing surgical residents for their abilities and aptitudes, to see who may be better equipped to gain operative skills with less training and coaching. Improving pedagogical techniques (Evans & Schenarts, 2016) may ease the difficulty on early learners, but the amount of training required for some individuals to achieve high surgical performance could still be prohibitive.

Techniques to measure performance limits continue to improve, prompted by increasing computational ability to quantify surgery in various contexts (Maier-Hein et al., 2017). There are increasing improvements automatically measuring performance in open (Azari et al., 2017; Mackenzie, Watts, Patel, Yang, Garofalo, et al., 2016) and laparoscopic procedures (Aggarwal et al., 2007), as well as with eye tracking (Richstone et al., 2010), and automatic “stroke” recognition (Ahmidi et al., 2015). The vast majority of these efforts, however, are limited to benchtop simulations or robot-assisted devices (Vedula et al., 2017).

1.4.2 Aptitude and Ability

Surgeons may lament a lack of manual dexterity and psycho-motor coordination (i.e. coordination, balance, haptic force control) of incoming residents, behaviors largely unpracticed in medical school until clinical rotations and outside the scope of common pre-medical undergraduate programs. In contrast, elite musicians and professional athletes often begin

deliberate practice in their field at an early age, engendering significant advantages later in life (Ericsson, Krampe, & Tesch-Römer, 1993). Decreased training time among residents (Nasca, Day, & Amis, 2010), little emphasis on mid-career training interventions (Bell, 2009; Cuschieri et al., 2001), and increasing complexity of the operating room (Bharathan, Aggarwal, & Darzi, 2013), have rekindled interest in "aptitude testing" to jumpstart selection and training of residents (Buckley et al., 2014; Moglia et al., 2014; Roitberg et al., 2013). Even "intellectual prowess" and "emotional stability" have been proposed as potential avenues to test for surgical aptitude (McDonald, 1998).

Aptitude is commonly described as a "natural" advantage a trainee brings to the table (Schendel, Shields, & Katz, 1974). This definition does not preclude new pedagogical techniques or better coaching from improving the performance of new trainees. Hislop et al. (2006), while examining aptitude for endovascular procedures, found that clinicians with extensive video game experience completed virtual reality tasks more quickly than those without prior video game experience. Willis et al. (2014), also found virtual reality and video game performances were related to one another, suggesting that pre-existing experience with video-games may transfer well to some simulations. Both studies sought to connect pre-existing strengths to an increased rate of performance gains during surgical training, relative to other trainees.

Existing literature, however, also tends to confound aptitude and ability. Ability is often qualified as "natural" "innate" or "fundamental," to describe an advantage someone brings to the selection process and training curriculum. Alfred Cuschieri (2003), for instance, uses aptitude as an intermediary to distinguish between skills (i.e. trained) and abilities (i.e. untrained): "abilities are the innate aptitudes that people can bring to given tasks and determine the level of proficiency that individuals obtain with training." In that view, skills require training, while abilities (being innate aptitudes) are brought

exclusively by the individual. Groenier et al. (2015) similarly uses “ability” to describe incoming trainee cognitive and psychomotor performance scores. In that study, participants with higher psychomotor scores learned to complete laparoscopic tasks more quickly, and with greater efficiency of movement. Szasz et al. (2016) also used “ability” to indicate a resident’s likelihood of promotion, measuring how they could meet performance thresholds for both the Objective Structured Assessment of Technical Skills (OSATS) and Objective Structured Assessment of Non-Technical Skills (OSANTS) concurrently. Moglia et al., (2014) refers to “innate aptitude” and “innate ability” synonymously in examining how psychomotor performance scores performance on the da Vinci Skills Simulator.

While individual strengths clearly impact outcomes of surgical training, merging aptitude and ability as a single concept creates problems defining performance. Reserving ability to express untrained strengths would suggest that no surgeon would have an ability to suture or complete any techniques where training is required. In contrast, ability is commonly used to describe training outcomes: Mattar et al. (2013) stipulates that many residents are “unable to operate for 30 unsupervised minutes of a major procedure” upon graduation. Referring to individually different strengths in absence of training or outside practice as abilities (whether fundamental, innate, or natural) limits the role of ability to account for trained skills.

To prevent overlapping interpretation between ability and aptitude, and to promote aptitude as a term in its own right, an individual’s pre-existing strength should be described as an aptitude that would impact the rate of performance gain. This would free the term “aptitude” from needing to rely on “ability” as a stepping stone. It allows “ability” to represent formally trained techniques that need not be innate or fundamental

and integrates aptitude into the training process. Separating these definitions maintains the notion of pre-existing strengths (i.e. aptitudes) without sacrificing the importance of training to increase surgical performance (i.e. abilities).

For this paper, ability represents the maximal performance an individual can offer under ideal circumstances, based on exigent training and previous experience. Recognizing that someone is able to perform a task implies they consistently meet or exceed some arbitrary threshold of acceptable performance (McGaghie, Miller, Sajid, & Telder, 1978). In this context, an ability would be demonstrable performance recognized as competent or higher. Ability grows over a career, commensurate with deliberate practice and exposure to difficult situations. Ability is also different than aptitude, as no trainee is considered "able" to perform a procedure because they score well on an aptitude test. However, both features are brought to bear in difficult operating room situations where surgeons use all advantages (trained or otherwise) to maximize performance.

Aptitude and ability are also domain-specific and nested within the taxonomy of Madani's framework discussed previously. For example, sewing aptitude and decision-making aptitude are quite different for early trainees, and ought to be described within the context of a relevant level of training and task, so as not to lose specificity or increase bias in selection. A software tool to provide quantitative feedback of performance without the need for coach intervention, could test for aptitudes, and help to improve student abilities before starting clinical rotations.

1.4.3 Experience

Lord Smith, a past President of The Royal College of Surgeons of England wrote "it would take me one year to teach a trainee how to do an operation, five years to teach them when

to do the operation, but a lifetime to teach them when not to do an operation" (2006). Experience describes the amount and breadth of familiarity a surgeon has in the operating room. This is typically expressed in the number of cases completed or amount of years in an attending role, but also manifests through "war stories" (Y. Y. Hu et al., 2012) and efficient "case scripts" or "illness scripts" (Norman, Eva, Brooks, & Hamstra, 2006; Schmidt et al., 1990) that improve expectations in the operating room.

Robust mental scripts show how experienced clinicians compose a litany of creative, fail-safe approaches to help deal with new challenges. These are complex representations that include "kinaesthetic and visual imagery" that help in psychomotor planning and movement (Holmes & Collins, 2001), and may be improved by mental practice (Louridas, Bonrath, Sinclair, Dedy, & Grantcharov, 2015). Previous experience helps form and represent critical patterns and cues which would otherwise go undetected. The set of expectations and scripts have also been described as a mental schema (Norman et al., 2006) to ease the burden of planning several steps ahead and recalling and integrating vast amounts of declarative knowledge on the fly (Sweller, 2008). By incorporating principles into patterns and schemas, surgeons also reduce the demand on cognitive resources, enabling greater flexibility to direct attention where most needed (O'Neil et al., 2014). Clinicians form crucial components of expertise by incorporating feature-based patterns and expectations of their own experience into these mental models (Schmidt et al., 1990).

In the search for objective measures, experienced surgeons offer a window into successful techniques that are tailored and honed over years of difficult practice. Carty and colleagues (2009) found that operative time decreased when surgeons reached 10 or more years of experience. Still, studying experienced surgeons – in the absence of more detailed objective measures – may fall short of guaranteeing positive outcomes. Geoff Norman summarized the

insufficiency of experience as a measure of surgical expertise in his personal correspondence to David Cook, facetiously suggesting that “gray hair and baldness would be good measures of expertise when comparing senior surgeons and third-year medical students” (Cook et al., 2014). Clarified by Norman’s wry attitude, experience is a necessary, albeit insufficient prerequisite for expertise. In fact, there is some evidence to support that individuals who consider themselves more experienced may be more resistant to contradictory information (Staats, KC, & Gino, 2018). Combating this trend, where it exists, and promoting so called “intellectual humility” (Gino, 2018) will be an ongoing effort.

1.4.4 Expertise

Understanding how surgical expertise develops is hindered by a lack of objective performance measures. This challenge is highlighted by Harald Mieg, who found that professionalism itself serves as the prevailing factor of expertise across fields where “standards of best practice still need to be established” (2009). Surgery fits the bill (Maier-Hein et al., 2017; Vedula et al., 2017), as clinicians “tend to conceptualize ‘mastery’ or ‘expertise’ as having conquered a specific set of skills, while other disciplines commonly associate these terms with a “continual learning state or perpetual devotion to improvement” (Greenberg & Klingensmith, 2015). Achieving expertise, rather than practicing it, reflects the strict social dichotomy between master and apprentice roles in surgical training.

Early attempts by Simon and Chase (1973) to define expertise emphasized direct testing of memory capacity. Although such measures provided a “convenient substitute for studies of actual performance” (Ericsson, 2005, p. 231), they did not provide sufficient explanations for the mechanisms supporting how expertise manifests across domains. Later approaches paralleled the proliferation of technology, describing the brain as a “computational device” which stored,

responded and retrieved information in “motor programs” analogous to software operating on computer platforms (Davids et al., 2008, p. 27).

The early emphasis on memory recall has been reframed over time to describe how expert performance manifests as a set of “exceptional skills in a particular domain” (Krawczyk, Bartlett, Kantarcioglu, Hamlen, & Thuraisingham, 2013, p. 364). Growing efforts to observe, measure and compare expert behavior (see Chi, 2011), have been bolstered by procedural and technological advances in cognitive task analysis (Tofel-Grehl & Feldon, 2013), brain imaging, (Krawczyk et al., 2013), protocol analysis and eye tracking (Charness & Tuffiash, 2016), simulation (Bond et al., 2008) and domain-specific factor analysis (Prietula, Feltovich, & Marchak, 2000), among others.

It is widely accepted today, that expertise is achieved through deliberate practice over time (Ericsson & Charness, 1994; Ericsson et al., 1993; Hashimoto et al., 2015; O’Neil et al., 2014; Palter & Grantcharov, 2014). Performers can target unique skills and reflect on their progress during planned periods where they “construct and seek out training situations in which the desired goal exceeds their current level of performance” (Ericsson, 2004). Exposure to difficult and variable situations is also critical (Spruit, Band, Hamming, & Ridderinkhof, 2014), enabling performers can become “adaptable for a range of varying performance characteristics” and “less vulnerable to transitory factors such as fatigue, audience effects, and anxiety” (Davids et al., 2008, p. 4). Unfortunately, the chance to reflect on these lessons is limited in the rapid and stressful environment of resident training (Jeffree & Clarke, 2010), resulting in a missed opportunity to promote metacognition and resident development (Pugh, 2014). Indeed, “practice without reflection and striving for continued improvement is a formula for mediocrity” (Weinbergger, Duffey, & Cassel, 2005).

Difficulty and expertise also have an intricate relationship. Experts "have adapted efficient ways to solve problems in their domains" and attack problems "by qualitatively different techniques" depending on the difficulty of the problem itself (Prietula, Feltovich, & Marchak, 2000, p. 64). Sufficient surgical training develops cue recognition, planning and error recovery during cases complex enough to integrate of both conscious (controlled) and unconscious (automated) actions (Pugh, Santacaterina, DaRosa, & Clark, 2011). An expert would thus be more responsive to the complexity of the task; able to conceptualize and plan a difficult operation at a high level of abstraction while integrating varying kinds of information (Ruis et al., 2017; van Merriënboer, Clark, & de Croock, 2002). Bond et al. (2008) supports this idea in describing how surgeons exhibiting superior performance use "pattern recognition to be efficient at the mundane," and "recognize when the pieces do not fit" (p. 1038), thereby adjusting their style of thinking to observe and adapt to evolving risks. Before the recent advances in developing surgical performance domains, recognizing these kinds of patterns, or "chunks" of knowledge, studied for years as part of successful performance in chess (Burns, 2004; Simon & Chase, 1973), was criticized for yielding "little direction for improving education of medical students" (Norman et al., 2006).

Jerome Groopman, author of *How Doctors Think* (2008), instead, frames the advantages afforded by medical expertise as a break from traditional training: "studies show that while it usually takes twenty to thirty minutes in a didactic exercise for the senior doctor and students to arrive at a working diagnosis, an expert clinician typically forms a notion of what is wrong with the patient within twenty seconds." He contends that practicing physicians pull cues in from all directions simultaneously – a non-linear mental process – but the taught method is rigidly linear: "Medical students are taught that the evaluation of a patient should proceed in a discrete linear way; you first take the patient's history, then perform a physical examination, order tests and

analyze the results. Only after all the data are compiled should you formulate hypotheses as to what might be wrong." In-context training for surgical residents is clearly crucial to build meaningful experience – despite the high cost to attending staff (Babineau et al., 2004).

Feltovich, Ford, & Hoffman (1997), creators of the TEMPEST model of expertise (Figure 2), describe expert adaptability in terms of preplanned actions. Termed "predictive encoding," the authors emphasize the role of advanced cognitive skills and command of dynamic knowledge in demonstrating expert performance. Such strategies help the expert draw on relevant experience, select useful tools and information, and balance the various forces at play during an operation.

The TEMPEST model highlights the various experiences, goals, materials and strategies of familiar tasks that experts may employ. Experts are driven and constrained by external forces like performance expectations, motives, and the rule of law. The "tail" – selection criteria, training, and professional standards – acts as a stabilizing force. The authors develop the model to represent completing a task.

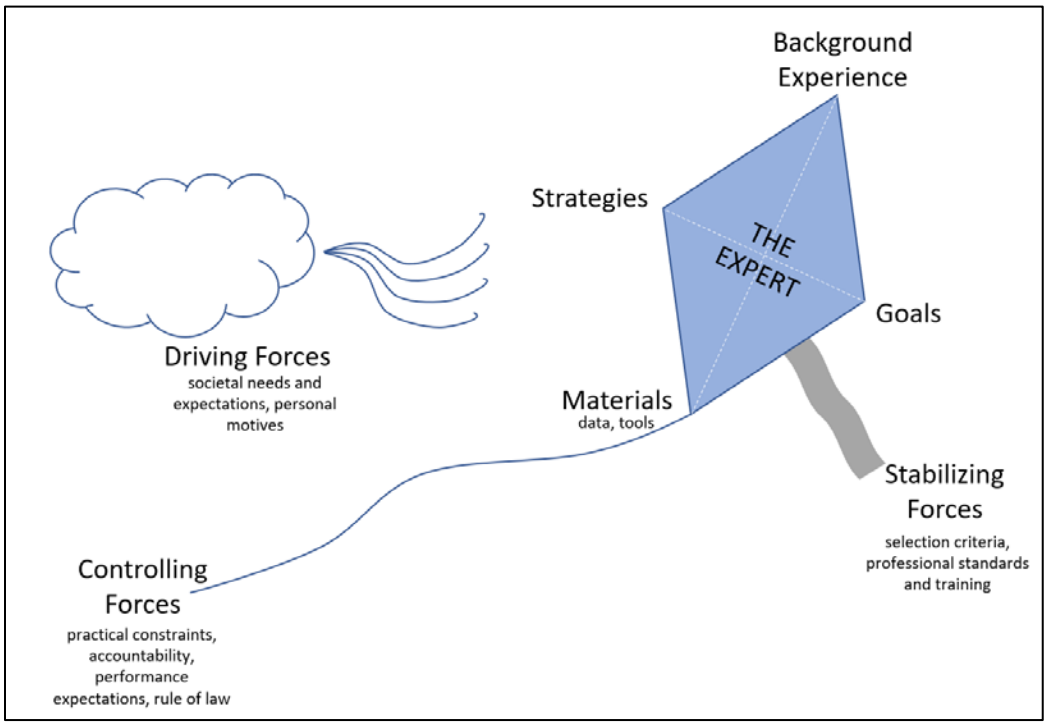


Figure 2: TEMPEST model describing the general framework of expertise (Adapted from Feltovich, Ford and Hoffman, 1997)

The TEMPEST model does not differentiate between the relative quality of performance between experts, even though the various inputs and forces imply performance among each expert are different. No two experts will have the exact same background experience or motivations. A helpful sorting scheme for this purpose is presented by Ericsson and colleagues (1993); arranging performers into “Least Accomplished”, “Good”, “Best” and “Professional,” but once again these boundaries lack objective thresholds. Surgeons, similarly, who might not pass an “Olympic” or “excellent” bar of performance, may be recognized as experts nonetheless (Alleman & Al-Assaf, 2005; Bell, 2009). Without a quantifiable performance standard, social standing and sense of professionalism serves as a stop-gap, with status conferred based on established case history, board certification, and leadership roles. Such attributes are highly valued and enshrined in the professional model of medicine, but social signifiers are only a piece of the puzzle in pursuing quantitative standards; they are notably absent in the TEMPEST model.

Expertise in surgery is thus challenging to pin down because it is used to refer both to the existing social hierarchy, and to the suite of practiced skills an individual can perform with consistency. In other words, surgical expertise continues to represent both professionalism and excellence; each of which are difficult, if not impossible, to define in their entirety. Professionalism is more easily understood; the Hippocratic oath is recognized centuries over as part of the social contract of medicine. Measures of "excellence" as Mieg describes, however, continue to evolve and face scrutiny as technology-assisted measurements of skill and performance grow (Vedula et al., 2017). These latter advances may be more effective than professionalism at promoting guidelines of physician best practices (see Laufer et al., 2016).

1.4.5 Competency

Competency, like expertise, is dual-faceted. A competent surgeon, and a competent surgical performance, for example, may describe different ideas. George Miller (1990) advocated that residents should achieve "competence," before performing and demonstrating skills on live patients. The US Accreditation Council for Graduate Medical Education (ACGME), meanwhile, is pushing for assessments across a series of job functions or "competencies," as part of the competency-based medical education (CBME) milestones project (ACGME, 2013). In light of these uses, a surgeon could be considered competent within a particular domain (i.e. consistently achieving a pre-defined rating in completing a procedure) or deemed competent overall (i.e. graduating from residency). For clarity, we define competency within a performance domain to mean meeting a quantifiable assessment threshold on a consistent and repeated basis. Referring to a clinician as a "competent surgeon," on the other hand, could also connote how that surgeon is perceived and trusted as a professional doctor, rather than how their skills have been quantifiably assessed. Firmly attaching competency to the underlying assessment content and

context will help to promote consistent interpretation of a clinician's practiced skills as performance assessments become increasingly embedded in surgical training.

The colloquial understanding of competency as something an individual ought to conquer, rather than practice, much like expertise, is a driving force in defining competence as a testable threshold of performance. These kinds of performance-based competency assessments are currently under development for laparoscopy (Miskovic et al., 2013) and for various skills associated with professionalism (Hochberg et al., 2010). Jelovsek et al., (2013) provides a broad overview of reliability evidence for operative assessments. For a discussion on evolving educational approaches, consult Evans and Schenarts (2016).

Outside of surgery, competency has been described as tantamount to the "attributes" arm of the Knowledge, Skills and Attributes (KSA) approach (McLagan, 1997). KSA is a lens popular for military analysis which frames attributes as task-applicable, but relatively domain independent and difficult to train. O'Neil, Perez, and Baker (2014) provide an in-depth discussion of the relationships between these constructs in their book *Teaching and Measuring Cognitive Readiness.*

1.4.6 Proficiency

Little emphasis is typically directed towards the difference between competency and proficiency. The Dreyfus Model of Skill Acquisition (Table 1, Dreyfuss & Dreyfus, 1980) offers a potential distinction, despite criticism for relying on intuition and omitting the utility of planning (Peña, 2010).

In the Dreyfus Model, competency is exhibited by active decision making and categorizing information. Proficiency, on the other hand, is analogous to Miller's "shows how" stage and reflects gains in operative autonomy as residents become attending surgeons. This

transition is earmarked by an increasing sense of responsibility commensurate with experience, and active demonstration of their abilities with decreasing oversight.

Table 1. The "Dreyfus Five Stage Model of Mental Stages in Skill Acquisition"(Dreyfuss & Dreyfus, 1980).

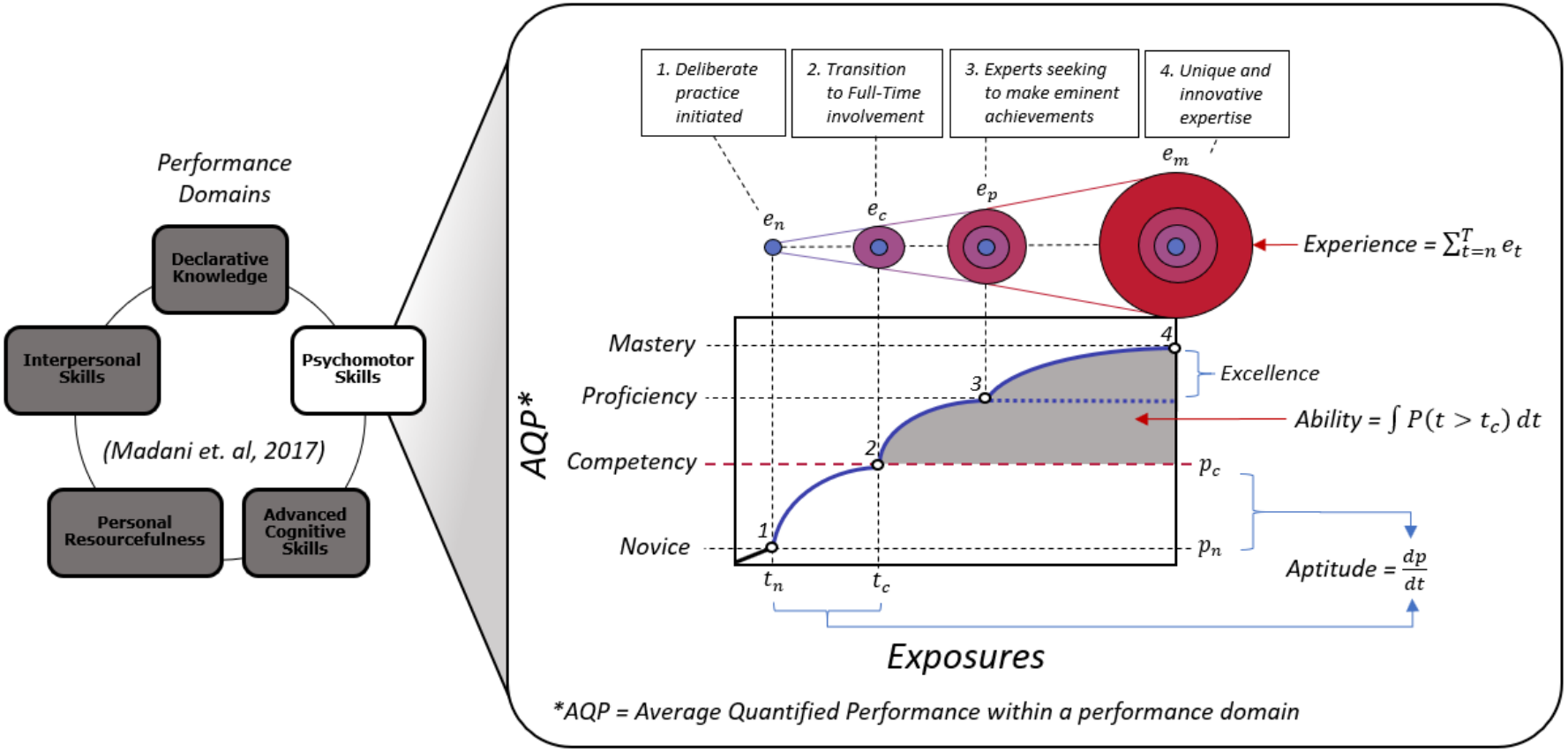
Stage	Autonomy	Mental Activities
1. Novice	Only feels responsible to follow the rules	Follows specific rules for specific situations. Rules are not conditional.
2. Advanced Beginner	Still does not experience personal responsibility	Begins to create and identify conditional rules. All decisions still follow rules.
3. Competent	Sense of responsibility arises from actively making decision	Learns organizing principles. Information sorting by relevance begins.
4. Proficient	Sense of responsibility increases with experience	Uses pattern recognition to assess what to do. Uses rules to determine how to do it.
5. Expert	Responsibility extends to others and the environment	No analysis or planning. Pattern recognition extends to plan as well as action.

Competency in the Dreyfus Model can also be construed as the lowest suitable level of performance. Proficiency, in contrast, represents greater consistency and responsibility, albeit not yet at levels considered "expert." In analyzing surgical skill, competency would represent meeting a minimum required assessment and starting to "actively make decisions" with autonomy. Brian George and colleagues (2014), in developing the Procedural Autonomy and Supervision System (PASS) on smartphones, describe similar stages of autonomy as "show and

tell," "active help", "passive help" and "supervision only." Proficiency represents consistency in performance in excess of competent levels. Proficiency would be characterized more by achieving repeated, stable, and efficient outcomes. Such thresholds would need to be drawn based on the procedure and task at hand and would revisited as disruptive technologies like laparoscopy are introduced.

1.5 Quantified Performance Model

The following model (Figure 3) is newly proposed to represent each of the skills-based terms discussed previously. The model represents performance as it develops over years of practice and exposure to challenging situations. Performance is represented as a quantifiable property over a surgeon's career and progresses through several stages, each with distinct exposure and responsibilities which change the rate of learning. Average performance is included to account for contextual and transitory factors. The model is adapted from the "three phases of development" model by Ericsson and Charness (1994). Competency is represented as an arbitrary, yet quantifiable threshold of performance and a gateway for increasing operative independence, with ability reserved to describe the sum of all actions for which a surgeon has already demonstrated competency – the integral under the continuous performance curve(s). Aptitude, on the other hand, is the rate (slope) of the curve from the time at which deliberate practice begins, to when a surgeon demonstrates competency on a regular basis. Experience, meanwhile, is represented as an expanding space and breadth of familiarity: the sum of unique exposure and instances of practice over the course of a career.



Note. Stages of performance curve from "Expert Performance: Its Structure and Acquisition" (Ericsson et. al, 1994). Average quantified performance would include valid measures of efficiency (e.g. fluidity, time to completion) and accuracy (e.g. successful completion).

Figure 3: Skills terminology model depicting relationship between common descriptors. Average Quantified Performance (AQP) is used to account for performance deviations due to transitory factors. The model combines both the Madani et al. (2017) performance domains framework, as well as the Three-Stage model of expert performance development, put forth by Ericsson and Charness (1994).

The Quantified Performance Model is designed to strike a balance between literature describing expertise and colloquial uses for the various terms at issue describing skill. Experience, for instance, while represented as a growing "bank" of lessons over the course of a career, is directly traceable to instances of deliberate practice and solving challenging cases as an attending. Excellent performance is superior to that of proficient and competent thresholds and associated with meaningful experiences. Because aptitude is typically used to describe the amount of effort or remedial training required at the beginning of surgical education, it is represented as the rate (slope) of the curve starting at the novice stage. The model assumes that the performance thresholds would be defined (and likely re-defined) as surgical technique and approach evolves – much as the current Milestones project defining various surgical competencies evolves (ACGME, 2013). Although several performance thresholds may be drawn, the thresholds for each of the performance stage (competency, proficiency, mastery) are intended to represent developments over the course of a surgical career.

Consider, as an example, the University of Wisconsin-Madison Urology Department residency training approach. In PGY 2, resident operations are "completely supervised by an attending faculty. The attention is on learning proper surgical skills, instrument identification and handling, and the proper steps to simple surgical procedures. By the completion of the [first] year, residents are expected to be able to perform all steps of simple surgical procedures with minimal guidance, but always under careful supervision."

Threshold performance at this stage is commensurate with the "novice" level. Operative autonomy is prohibited, and aptitude for general surgical practices is still being evaluated. Deliberate practice of in-vivo surgical skill is minimal but grows over the course of the year; initializing a set of experiences from which to form basic "illness scripts." At this stage, it may

be reasonable to use performance assessments to generate formative feedback, or to examine aptitude.

Between the third (PGY 3) and fourth (PGY 4) years, residents are expected to move up the performance curve. There are “increasing opportunities to conduct certain steps” as skills develop. Residents approach “conducting an entire procedure independently,” albeit under direct supervision. Mentors provide “immediate feedback and remediation of any deficiencies” (PGY 4). Within the Quantified Performance Model, these improvements describe progress towards competent levels of performance in each domain. Instances of deliberate practice are increasing and building experience commensurate with the “active help” stage of PASS.

By the time a resident enters their 5th year, and as a chief resident, they are expected to “perform all steps of major urologic surgeries,” and achieve “autonomy in performing basic surgical procedures.” At this point, residents working with autonomy are not considered or consulted as experts, but their experience enables greater independence in the operating room and helps to inform those of lesser training. Objective assessments within the Quantified Performance Model may indicate a competent level in many of the performance domains for several procedures. Only after consistent passage, however, would that resident be considered “competent” for those procedures. Residents may also exhibit proficient levels of performance for a handful of simpler bedside or out-patient procedures. Completion of these five PGYs would be similar to completing the first three stages in the Dreyfus model of skill acquisition.

As residents become attending surgeons and transition to the new expectations of their full-time role, they would pursue a proficient level of performance. In contrast to residency, which places greater emphasis on focusing attention on to enable periods of deliberate practice, time spent operating as an attending surgeon places greater emphasis on achieving efficient, positive outcomes, even in difficult situations. One study found that variation in operation time

and complication rates during mammaplasty stabilized only after 12 years of active practice as an attending surgeon (Carty et al., 2009). Still, lapses in planning, neglecting to complete or recall steps in an operation, or increased variability may indicate declines in performance at any stage in a surgeon's career. To enable consistent results in the OR, these surgeons would need exhibit more efficient and robust error-management techniques. It is reasonable to assume that an attending performing at a master level have reached this point of stability and gained 12 or more years of experience. Within the proposed model, performance traits of an individual exhibiting mastery would serve as a template and resource to improve the rate of achieving proficiency for others in various difficult procedures.

As a surgeon progresses through these stages and strives for higher levels of performance, assessments would need to target more complex attributes of Madani's domains. Proficiency would need to be assessed through clinical simulations of increasingly difficult scenarios. At the same time, however, testing of previously surpassed performance thresholds would expose areas of needed practice to maintain skill with age, changes in life circumstance, or to demonstrate readiness to transition to another kind of surgery. Quantified performance testing throughout a surgical career may also serve to share expert strategies and mental models, while limiting patient exposure. Identifying features of performance commensurate with advanced tenure (i.e. expert and master surgeons) is an ongoing avenue of research.

1.6 Discussion

The standard lexicon proposed by the Quantified Performance Model (QPM) of surgical skills terminology (Figure 3) focuses on defining performance as surgeons gain skills, age, and eventually retire. It is applicable to each of the five surgical performance domains (Madani et al., 2017) and is particularly timely for increasing efforts to quantify psychomotor skills. The model incorporates the role of deliberate practice in building expertise and paves the way to frame

operative assessments as a consistent, repeated demonstration of performance rather than a one-time credential. The model does not, on the other hand, specify the content of these assessments. Described as an “instructional design problem,” developing meaningful assessments is a continuing area of research subject to validity (Kane, 2006) and overall utility (van der Vleuten & Schuwirth, 2005) analysis.

It is assumed that assessments will continue to adapt as technology and surgical techniques evolve. Procedure difficulty must also be considered. Planning, situational awareness, or other “advanced cognitive skills,” for instance, may expose greater abilities in experienced clinicians than less complex assessments of salient psychomotor skills while suturing on simulated benchtop models.

The proposed model integrates potential assessment measures as an attempt to reach competency, and as a building block to proficiency (much like the Dreyfus model). Competency represents a transition to increasing responsibility and operative autonomy. The model reflects George Miller’s focus on being able to show or perform skills. In addition, it uniquely frames performance as a repeated and consistent measure, to account for situational context and variation. It supports regular, repeated performance testing and reflects the ongoing push to demonstrate skills over time, even as they degrade due to advanced age or change in professional status. The model also supports the construction of various pass-fail thresholds, fitting well within the rhetoric of Madani, by encouraging active assessment for “a competent level of performance” (2017).

Educational literature often uses similar terms, however, to detail a pedagogical approach. Consistently reaching a competent threshold, for example, is commonly referred to as “mastery” of the assessment topic (McGaghie et al., 1978). In contrast, “master” surgeons are commonly described as those with substantial operative experience and involved in training

efforts (Cuschieri et al., 2001). Performance, too, has held a unique educational meaning, connoting the final stage of one-on-one manual skill training (Peyton, 1998). Over time, the term "execution" has replaced "performance" in the these contexts (Munster, Stosch, Hindrichs, Franklin, & Matthes, 2016), with performance describing a more continuous scale of development (Jeffree & Clarke, 2010) similar to the proposed model.

Over-simplifying any assessment framework poses a natural challenge to physician and patient autonomy – an evolving, yet fundamental tenet in the professional model of medicine (ABIM Foundation, 2002). If quantification of skill is implemented poorly and becomes anathema to the "secret glory" of medicine as a craft profession (Donabedian, 1988), surgeons may opt instead to retreat to their respective corners; offering additional challenges to the already difficult prospect of competency-based medical education (Touchie & Ten Cate, 2016). Worthwhile assessments could be overlooked before they have a chance to mature – undermining improvements to quality patient care and wasting valuable resources. Graham and Deary (1991) argued that widespread adoption of such testing requires maturity of three things: robust understanding of skill, studies with subjective ratings as dependent variables, and an appropriate "working definition of superior surgical performance."

The proposed model in this paper offers a productive and traceable way to use surgical skills terminology in quantifying performance. The model integrates well with existing validity frameworks by promoting clear inferences and uses throughout a surgical career. To promote easier adoption among the medical community, the proposed definitions integrate existing literature and colloquial understanding.

1.7 Conclusion

This paper has focused on defining surgical skill terms that, despite their ubiquity, lack unique interpretations. A novel model of terminology is proposed to assist in framing objective

and feature-based surgical skills along a continuous scale of performance. Experience is represented as a growing “bank” of exposure to difficult situations; and includes the sum of instances of deliberate practice. Competency represents an arbitrary performance threshold, generally commensurate with graduation from residency and full-time involvement as an attending surgeon. Proficiency is characterized by decreasing variation and increasingly efficient outcomes. While attendings will pursue proficiency for the most difficult and complex operations, some residents may also reach proficient levels of performance for familiar operations and bedside procedures. Ability represents all performance a surgeon can offer in excess of a competent level, drawn as the integral under the performance curve after reaching competency. Aptitude is the rate at which one could achieve a competent performance level, given current pedagogical techniques. Mastery represents a performance threshold in excess of proficiency; characterized by excellent outcomes and novel techniques beyond those expected at proficient levels. Descriptors like elite and superior may be reserved for performances at the mastery level.

Many of these definitions (consider competence, for example) depend on reaching a quantitative threshold of performance that has yet to reach maturity. Establishing validity evidence for such assessments in accordance with modern frameworks (Kane, 2013) is ongoing. The proposed model frames quantitative assessments within a continuous performance curve throughout stages of a surgical career. Each stage is associated with different training regiments and responsibilities, adapted from the “three phases of development” model by Ericsson and Charness (1994). To be considered competent or proficient to conduct an operation, a surgeon would need to consistently and repeatedly meet those relevant performance thresholds for relevant assessments in each surgical domain (Madani et al., 2017).

As objective surgical skill analysis research continues to grow, consistent terminology will be critical in translating objective measures into formative feedback, and eventually, valid assessments. The quantified performance model – accompanied by increasing abilities to measure performance – may aid in clarifying the duality of surgical expertise as a measure of professionalism and excellence. It may never be possible to quantify the artistry inherent in advanced surgery or define unique attributes of skill for complex operations. But, it may be possible to identify performance with enough specificity to discern surgeon progression from novice, to competent, proficient, and beyond. These thresholds could facilitate training, aptitude testing, placement, remediation, and timing of professional transition or retirement.

1.8 References

ABIM Foundation. (2002). Medical professionalism in the new millennium: a physician charter. Annals of Internal Medicine, 136(3), 243–6. http://doi.org/10.7326/0003-4819-136-3-200202050-00012

ACGME. (2013). Accrediation Council for Graduate Medical Eduction (ACGME). Retrieved May 1, 2018, from http://www.acgme.org/What-We-Do/Accreditation/Milestones/

Aggarwal, R., Grantcharov, T., Moorthy, K., Milland, T., Papasavas, P., Dosis, A., … Darzi, A. (2007). An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Annals of Surgery, 245(6), 992–999. http://doi.org/10.1097/01.sla.0000262780.17950.e5

Ahmidi, N., Poddar, P., Jones, J. D., Vedula, S. S., Ishii, L., Hager, G. D., & Ishii, M. (2015). Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. International Journal of Computer Assisted Radiology and Surgery, 10(6), 981–991. http://doi.org/10.1007/s11548-015-1194-1

Alleman, A. M., & Al-Assaf, A. F. (2005). Have You Wondered About Your Colleague's Surgical Skills? American Journal of Medical Quality, 20(2), 78–82. http://doi.org/10.1177/1062860604273746

Anderson, J. R. (1982). Acquisition of Cognitive Skill. Psychological Review, 89(4), 369–406. http://doi.org/10.1037/0033-295X.89.4.369

Azari, D. P., Frasier, L. L., Quamme, S. R. P., Greenberg, C. C., Pugh, C. M., Greenberg, J. A., & Radwin, R. G. (2017). Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. Annals of Surgery, XX(Xx), 1. http://doi.org/10.1097/SLA.0000000000002478

Azari, D. P., Pugh, C. M., Laufer, S., Kwan, C., Chen, C. H., Yen, T. Y., … Radwin, R. G. (2015). Evaluation of Simulated Clinical Breast Exam Motion Patterns Using Marker-Less Video Tracking. Human Factors, 58(3), 427–440. http://doi.org/10.1177/0018720815613919

Babineau, T. J., Becker, J., Gibbons, G., Sentovich, S., Hess, D., Robertson, S., & Stone, M. (2004). The "cost" of operative training for surgical residents. Archives of Surgery (Chicago, Ill. : 1960), 139(4), 366-9; discussion 369-70. http://doi.org/10.1001/archsurg.139.4.366

Bell, R. H. (2009). Why Johnny cannot operate. Surgery, 146(4), 533–542. http://doi.org/10.1016/j.surg.2009.06.044

Bharathan, R., Aggarwal, R., & Darzi, A. (2013). Operating room of the future. Best Practice and Research: Clinical Obstetrics and Gynaecology, 27(3), 311–322. http://doi.org/10.1016/j.bpobgyn.2012.11.003

Bond, W., Kuhn, G., Binstadt, E., Quirk, M., Wu, T., Tews, M., … Ericsson, K. A. (2008). The Use of Simulation in the Development of Individual Cognitive Expertise in Emergency Medicine. Academic Emergency Medicine, 15(11), 1037–1045. http://doi.org/10.1111/j.1553-2712.2008.00229.x

Buckley, C. E., Kavanagh, D. O., Nugent, E., Ryan, D., Traynor, O. J., & Neary, P. C. (2014). The impact of aptitude on the learning curve for laparoscopic suturing. American Journal of Surgery, 207(2), 263–270. http://doi.org/10.1016/j.amjsurg.2013.08.037

Burns, B. D. (2004). The effects of speed on skilled chess performance. Psychological Science, 15(7), 442–447. http://doi.org/10.1111/j.0956-7976.2004.00699.x

Carty, M. J., Chan, R., Huckman, R., Snow, D., & Orgill, D. P. (2009). A detailed analysis of the reduction mammaplasty learning curve: A statistical process model for approaching surgical performance improvement. Plastic and Reconstructive Surgery, 124(3), 706–714. http://doi.org/10.1097/PRS.0b013e3181b17a13

Charness, N., & Tuffiash, M. (2016). The Role of Expertise Research and Human Factors in Capturing , Explaining , and Producing Superior Performance, 50(3), 427–432. http://doi.org/10.1518/001872008X312206.

Chi, M. T. H. (2011). Theoretical Perspectives, Methodological Appraoches, and Trends in the Study of Expertise. Expertise in Mathematics Instruction, 17–39. http://doi.org/10.1007/978-1-4419-7707-6

Choudhry, N. K., Fletcher, R. H., & Soumerai, S. B. (2005). Improving Patient Care Systematic Review : The Relationship between Clinical Experience and Quality of Health Care.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. Medical Education, 49(6), 560–575. http://doi.org/10.1111/medu.12678

Cook, D. A., & Reed, D. A. (2015). Appraising the Quality of Medical Education Research Methods. Academic Medicine, 90(8), 1067–1076. http://doi.org/10.1097/ACM.0000000000000786

Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Advances in Health Sciences Education, 19(2), 233–250. http://doi.org/10.1007/s10459-013-9458-4

Cuschieri, A., Francis, N., Crosby, J., & Hanna, G. B. (2001). What do master surgeons think of surgical competence and revalidation? American Journal of Surgery, 182(2), 110–116. http://doi.org/10.1016/S0002-9610(01)00667-5

Datta, V., Chang, A., Mackay, S., & Darzi, A. (2002). The relationship between motion analysis and surgical technical assessments. The American Journal of Surgery, 184(1), 70–73. http://doi.org/10.1016/S0002-9610(02)00891-7

Davids, K., Button, C., & Bennett, S. (2008). Dynamics of Skill Acquition - A Constraints-Led Approach. Champaign, IL, USA: Human Kinetics.

Dedy, N. J., Fecso, A. B., Szasz, P., Bonrath, E. M., & Grantcharov, T. P. (2016). Implementation of an Effective Strategy for Teaching Nontechnical Skills in the Operating Room. Annals of Surgery, 263(5), 937–941. http://doi.org/10.1097/SLA.0000000000001297

Donabedian, A. (1988). The quality of care. How can it be assessed? JAMA: The Journal of the American Medical Association, 260(12), 1743–1748. http://doi.org/10.1001/jama.260.12.1743

Dreyfuss, S. E., & Dreyfus, H. L. (1980). A five-stage model of the mental activities involved in directed skill acquisition. Operations Research Center, (February), 1–18. http://doi.org/ADA084551

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Academic Medicine : Journal of the Association of American Medical Colleges, 79(10 Suppl), S70–S81. http://doi.org/10.1097/00001888-200410001-00022

Ericsson, K. A. (2005). Recent advances in expertise research: A commentary on the contributions to the special issue. Applied Cognitive Psychology, 19(2), 233–241. http://doi.org/10.1002/acp.1111

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. American Psychologist, 49(8), 725–747. http://doi.org/10.1037/0003-066X.49.8.725

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100(3), 363–406. http://doi.org/10.1037/0033-295X.100.3.363

Evans, C. H., & Schenarts, K. D. (2016). Evolving Educational Techniques in Surgical Training. Surgical Clinics of North America, 96(1), 71–88. http://doi.org/10.1016/j.suc.2015.09.005

Feltovich, P. J., Ford, K. M., & Hoffman, R. R. (1997). Exertise in Context. Cambridge: The MIT Press.

Fitts, P. M., & Posner, M. L. I. (1967). Human Performance. Belmont, CA: Brooks/Cole Publishing Co; Retrieved from http://www.worldcat.org/title/human-performance/oclc/00480476

Francis, D. M. A. (2009). Surgical decision making. ANZ Journal of Surgery, 79(12), 886–891. http://doi.org/10.1111/j.1445-2197.2009.05139.x

Frasier, L. L., Azari, D. P., Ma, Y., Quamme, S. R. P., Radwin, R. G., Pugh, C. M., … Greenberg, C. C. (2016). A marker-less technique for measuring kinematics in the operating room. Surgery. http://doi.org/10.1016/j.surg.2016.05.004

George, B. C., Teitelbaum, E. N., Meyerson, S. L., Schuller, M. C., Darosa, D. A., Petrusa, E. R., … Fryer, J. P. (2014). Reliability, validity, and feasibility of the zwisch scale for the assessment of intraoperative performance. Journal of Surgical Education, 71(6), e90–e96. http://doi.org/10.1016/j.jsurg.2014.06.018

Gino, F. (2018, April). Why Too Much Experience Can Backfire. Scientific American. Retrieved from https://www.scientificamerican.com/article/why-too-much-experience-can-backfire/

Ginsburg, S., Regehr, G., & Lingard, L. (2004). Basing the evaluation of professionalism on observable behaviors: a cautionary tale. Academic Medicine : Journal of the Association of American Medical Colleges, 79(10 Suppl), S1-4. http://doi.org/10.1097/00001888-200410001-00001

Glarner, C. E., Hu, Y. Y., Chen, C. H., Radwin, R. G., Zhao, Q., Craven, M. W., … Greenberg, C. C. (2014). Quantifying technical skills during open operations using video-based motion analysis. Surgery (United States), 156(3), 729–734. http://doi.org/10.1016/j.surg.2014.04.054

Graham, K. S., & Deary, I. J. (1991). A role for aptitude testing in surgery? Journal of the Royal College of Surgeons of Edinburgh, 36(2), 70–4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2051420

Greenberg, C. C., Ghousseini, H. N., Pavuluri Quamme, S. R., Beasley, H. L., & Wiegmann, D. A. (2015). Surgical Coaching for Individual Performance Improvement. Annals of Surgery, 261(1), 32–34. http://doi.org/10.1097/SLA.0000000000000776

Greenberg, C. C., & Klingensmith, M. E. (2015). The continuum of coaching: Opportunities for surgical improvement at all levels. Annals of Surgery, 262(2), 217–219. http://doi.org/10.1097/SLA.0000000000001290

Groenier, M., Groenier, K. H., Miedema, H. A. T., & Broeders, I. A. M. J. (2015). Perceptual Speed and Psychomotor Ability Predict Laparoscopic Skill Acquisition on a Simulator.

Journal of Surgical Education, 72(6), 1224–1232. http://doi.org/10.1016/j.jsurg.2015.07.006

Groopman, J. (2008). How Doctors Think. Mariner Books.

Hashimoto, D. A., Sirimanna, P., Gomez, E. D., Beyer-Berjot, L., Ericsson, K. A., Williams, N. N., … Aggarwal, R. (2015). Deliberate practice enhances quality of laparoscopic surgical performance in a randomized controlled trial: from arrested development to expert performance. Surgical Endoscopy, 29(11), 3154–3162. http://doi.org/10.1007/s00464-014-4042-4

Hislop, S. J., Hsu, J. H., Narins, C. R., Gillespie, B. T., Jain, R. A., Schippert, D. W., … Illig, K. A. (2006). Simulator assessment of innate endovascular aptitude versus empirically correct performance. Journal of Vascular Surgery, 43(1), 47–55. http://doi.org/10.1016/j.jvs.2005.09.035

Hochberg, M. S., Kalet, A., Zabar, S., Kachur, E., Gillespie, C., & Berman, R. S. (2010). Can professionalism be taught? Encouraging evidence. American Journal of Surgery, 199(1), 86–93. http://doi.org/10.1016/j.amjsurg.2009.10.002

Holmes, P. S., & Collins, D. J. (2001). The PETTLEP Approach to Motor Imagery: A Functional Equivalence Model for Sport Psychologists. Journal of Applied Sport Psychology, 13(1), 60–83. http://doi.org/10.1080/10413200109339004

Hu, Y. Y., Peyre, S. E., Arriaga, A. F., Roth, E. M., Corso, K. A., & Greenberg, C. C. (2012). War stories: A qualitative analysis of narrative teaching strategies in the operating room. American Journal of Surgery, 203(1), 63–68. http://doi.org/10.1016/j.amjsurg.2011.08.005

Hull, L., Arora, S., Aggarwal, R., Darzi, A., Vincent, C., & Sevdalis, N. (2012). The impact of nontechnical skills on technical performance in surgery: A systematic review. Journal of the American College of Surgeons, 214(2), 214–230. http://doi.org/10.1016/j.jamcollsurg.2011.10.016

Jeffree, R. L., & Clarke, R. M. (2010). Ten tips for teaching in the theatre tearoom: Shifting the focus from teaching to learning. World Journal of Surgery, 34(11), 2518–2523. http://doi.org/10.1007/s00268-010-0719-6

Jelovsek, J. E., Kow, N., & Diwadkar, G. B. (2013). Tools for the direct observation and assessment of psychomotor skills in medical trainees: A systematic review. Medical Education, 47(7), 650–673. http://doi.org/10.1111/medu.12220

Kane, M. (2006). Validation. In R. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). The Argument-Based Approach to Validation. School Psychology Review, 42(4), 448–457.

Kelly, J. F., Ritenour, A. E., McLaughlin, D. F., Bagg, K. a, Apodaca, A. N., Mallak, C. T., … Holcomb, J. B. (2008). Injury Severity and Causes of Death From Operation Iraqi

Freedom and Operation Enduring Freedom: 2003–2004 Versus 2006. Journal of Trauma and Acute Care Surgery, 64(2), S21-S26; discussion S26-S27. http://doi.org/10.1097/TA.0b013e318160b9fb

Krawczyk, D., Bartlett, J., Kantarcioglu, M., Hamlen, K., & Thuraisingham, B. (2013). Measuring expertise and bias in cyber security using cognitive and neuroscience approaches. In 2013 IEEE International Conference on Intelligence and Security Informatics (pp. 364–367). IEEE. http://doi.org/10.1109/ISI.2013.6578859

Laufer, S., D'Angelo, A.-L. D., Kwan, C., Ray, R. D., Yudkowsky, R., Boulet, J. R., … Pugh, C. M. (2016). Rescuing the Clinical Breast Examination. Annals of Surgery, XX(X), 1–6. http://doi.org/10.1097/SLA.0000000000002024

Law Forsyth, K., DiMarco, S. M., Jenewein, C. G., Ray, R. D., D'Angelo, A.-L. D., Cohen, E. R., … Pugh, C. M. (2017). Do errors and critical events relate to hernia repair outcomes? The American Journal of Surgery, 213(4), 652–655. http://doi.org/10.1016/j.amjsurg.2016.11.020

Louridas, M., Bonrath, E. M., Sinclair, D. A., Dedy, N. J., & Grantcharov, T. P. (2015). Randomized clinical trial to evaluate mental practice in enhancing advanced laparoscopic surgical performance. British Journal of Surgery, 102(1), 37–44. http://doi.org/10.1002/bjs.9657

Mackenzie, C. F., Watts, D., Patel, R. R., Yang, S., Garofalo, E., Puche, A. C., … Tisherman, S. A. (2016). Sensor-Free Computer Vision Hand-Motion Entropy and Video Analysis of Technical Performance During Open Vascular Surgery: Proof of Concept for Methodology. Journal of the American College of Surgeons, 223(4), e63. http://doi.org/10.1016/j.jamcollsurg.2016.08.166

Madani, A., Vassiliou, M. C., Watanabe, Y., Al-Halabi, B., Al-Rowais, M. S., Deckelbaum, D. L., … Feldman, L. S. (2017). What Are the Principles That Guide Behaviors in the Operating Room? Annals of Surgery, 265(2), 255–267. http://doi.org/10.1097/SLA.0000000000001962

Maier-Hein, L., Vedula, S., Speidel, S., Navab, N., Kikinis, R., Park, A., … Jannin, P. (2017). Surgical Data Science: Enabling Next-Generation Surgery. Retrieved from https://arxiv.org/abs/1701.06482

McDonald, P. (1998). Training for surgeons after the year 2000. J. R. Soc. Med., 91(8), 401.

McGaghie, W. C., Miller, G. E., Sajid, A. W., & Telder, T. W. (1978). Competency-Based Curriculum Development in Medical Education: An Introduction. Public Health Papers. Retrieved from http://eric.ed.gov/?id=ED168447%5Cnhttp://apps.who.int/iris/handle/10665/39703%5Cnhttp://whqlibdoc.who.int/php/WHO_PHP_68.pdf

McLagan, P. a. (1997). Competencies: the next generation. Training & Development.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. Academic Medicine. http://doi.org/10.1097/00001888-199009000-00045

Miskovic, D., Ni, M., Wyles, S. M., Kennedy, R. H., Francis, N. K., Parvaiz, A., … Hanna, G. B. (2013). Is Competency Assessment at the Specialist Level Achievable? A Study for the National Training Programme in Laparoscopic Colorectal Surgery in England. Annals of Surgery, 257(3), 476–482. http://doi.org/10.1097/SLA.0b013e318275b72a

Moglia, A., Ferrari, V., Morelli, L., Melfi, F., Ferrari, M., Mosca, F., & Cuschieri, A. (2014). Distribution of innate ability for surgery amongst medical students assessed by an advanced virtual reality surgical simulator. Surgical Endoscopy and Other Interventional Techniques, 28(6), 1830–1837. http://doi.org/10.1007/s00464-013-3393-6

Moorthy, K., Munz, Y., Adams, S., Pandey, V., & Darzi, A. (2005). A Human Factors Analysis of Technical and Team Skills Among Surgical Trainees During Procedural Simulations in a Simulated Operating Theatre. Annals of Surgery, 242(5), 631–639. http://doi.org/10.1097/01.sla.0000186298.79308.a8

Munster, T., Stosch, C., Hindrichs, N., Franklin, J., & Matthes, J. (2016). Peyton's 4-Steps-Approach in comparison: Medium-term effects on learning external chest compression - a pilot study. GMS Journal for Medical Education, 33(4), Doc60. http://doi.org/10.3205/zma001059

Murinson, B. B., Agarwal, A. K., & Haythornthwaite, J. A. (2008). Cognitive Expertise, Emotional Development, and Reflective Capacity: Clinical Skills for Improved Pain Care. Journal of Pain, 9(11), 975–983. http://doi.org/10.1016/j.jpain.2008.07.010

Nasca, T. J., Day, S. H., & Amis, E. S. (2010). The New Recommendations on Duty Hours from the ACGME Task Force. New England Journal of Medicine, 363(2), e3. http://doi.org/10.1056/NEJMsb1005800

Nathwani, J. N., Glarner, C. E., Law, K. E., McDonald, R. J., Zelenski, A. B., Greenberg, J. A., & Foley, E. F. (2017). Integrating Postoperative Feedback Into Workflow: Perceived Practices and Barriers. Journal of Surgical Education, 74(3), 406–414. http://doi.org/10.1016/j.jsurg.2016.11.001

Norman, G., Eva, K., Brooks, L., & Hamstra, S. (2006). Expertise in Medicine and Surgery. The Cambridge Handbook of Expertise and Expert Performance, 339–54.

O'Neil, H. F., Perez, R. S., & Baker, E. L. (2014). Teaching and Measuring Cognitive Readiness. (H. F. O'Neil, R. S. Perez, & E. L. Baker, Eds.)Teaching and Measuring Cognitive Readiness (Vol. 9781461475). Boston, MA: Springer US. http://doi.org/10.1007/978-1-4614-7579-8

Palter, V. N., & Grantcharov, T. P. (2014). Individualized deliberate practice on a virtual reality simulator improves technical performance of surgical novices in the operating room: a randomized controlled trial. Annals of Surgery, 259(3), 443–8. http://doi.org/10.1097/SLA.0000000000000254

Peña, A. (2010). The Dreyfus model of clinical problem-solving skills acquisition: a critical perspective. Medical Education Online, 15, 1–11. http://doi.org/10.3402/meo.v15i0.4846

Perdanasari, A. T., & Hollier, L. H. (2017). Review of "What Are the Principles That Guide Behaviors in the Operating Room? Creating a Framework to Define and Measure Performance" by Madani A, Vassiliou MC, Watanabe Y, Al-Halabi B, Al-Rowais MS, Deckelbaum DL, Fried GM, Feldman LS in Ann Surg 265. Journal of Craniofacial Surgery, 28(3), 842. http://doi.org/10.1097/SCS.0000000000003424

Perez, R. S., Skinner, A., Weyhrauch, P., Niehaus, J., Lathan, C., Schwaitzberg, S. D., & Cao, C. G. L. (2013). Prevention of surgical skill decay. Military Medicine, 178(10 Suppl), 76–86. http://doi.org/10.7205/MILMED-D-13-00216

Peyton, W. M. (1998). Teaching in the Theathre. Teaching and Learning in Medical Practice, (S.), 171–180.

Prietula, M. J., Feltovich, P. J., & Marchak, F. (2000). Factors influencing analysis of complex cognitive tasks: a framework and example from industrial process control. Human Factors, 42(1), 56–74. http://doi.org/10.1518/001872000779656589

Pugh, C. M. (2014). Getting a sense for the surgical touch [Video File]. Retrieved from https://www.tedmed.com/talks/show?id=292997

Pugh, C. M., & DaRosa, D. a. (2013). Use of cognitive task analysis to guide the development of performance-based assessments for intraoperative decision making. Military Medicine, 178(10 Suppl), 22–7. http://doi.org/10.7205/MILMED-D-13-00207

Pugh, C. M., Santacaterina, S., DaRosa, D. A., & Clark, R. E. (2011). Intra-operative decision making: More than meets the eye. Journal of Biomedical Informatics, 44(3), 486–496. http://doi.org/10.1016/j.jbi.2010.01.001

Radwin, R. G., Azari, D. P., Frasier, L., Quamme, S. R. P., Chen, C.-H., Yen, T., … Greenberg, C. C. (2014). A Marker-less Video Tracking Approach for Quantifying Open Surgical Skill. In International Annual Meeting of the Human Factors and Ergonomics Society: Panel Session (pp. 924–928). Panel Session.

Regenbogen, S. E., Greenberg, C. C., Studdert, D. M., Lipsitz, S. R., Zinner, M. J., & Gawande, A. a. (2007). Patterns of Technical Error Among Surgical Malpractice Claims. Annals of Surgery, 246(5), 705–711. http://doi.org/10.1097/SLA.0b013e31815865f8

Reznick, R. K., & MacRae, H. (2006). Teaching Surgical Skills — Changes in the Wind. New England Journal of Medicine, 355(25), 2664–2669. http://doi.org/10.1056/NEJMra054785

Richstone, L., Schwartz, M. J., Seideman, C., Cadeddu, J., Marshall, S., & Kavoussi, L. R. (2010). Eye metrics as an objective assessment of surgical skill. Annals of Surgery, 252(1), 177–182. http://doi.org/10.1097/SLA.0b013e3181e464fb

Rogers, S. O., Gawande, A. A., Kwaan, M., Puopolo, A. L., Yoon, C., Brennan, T. A., & Studdert, D. M. (2006). Analysis of surgical errors in closed malpractice claims at 4 liability insurers. Surgery, 140(1), 25–33. http://doi.org/10.1016/j.surg.2006.01.008

Roitberg, B., Banerjee, P., Luciano, C., Matulyauskas, M., Rizzi, S., Kania, P., & Gasco, J. (2013). Sensory and motor skill testing in neurosurgery applicants: A pilot study using a virtual reality haptic neurosurgical simulator. Neurosurgery, 73(SUPPL. 4), 116–121. http://doi.org/10.1227/NEU.0000000000000089

Ruis, A. R., Rosser, A. A., Quandt-Walle, C., Nathwani, J. N., Shaffer, D. W., & Pugh, C. M. (2017). The Hands and Head of a Surgeon: Modeling Operative Competency with Multimodal Epistemic Network Analysis. The American Journal of Surgery. http://doi.org/10.1016/j.amjsurg.2017.11.027

Rutherford, D. N., D'Angelo, A.-L. D., Law, K. E., & Pugh, C. M. (2015). Advanced Engineering Technology for Measuring Performance. Surgical Clinics of North America, 95(4), 813–826. http://doi.org/10.1016/j.suc.2015.04.005

Schendel, J. D., Shields, J. L., & Katz, M. G. (1974). Retention of motor skills: Review. Technical Paper, 50.

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. (1990). A cognitive perspective on medical expertise: theory and implication. Academic Medicine : Journal of the Association of American Medical Colleges, 65(10), 611–21. http://doi.org/10.1097/00001888-199010000-00001

Scott, D. J., Cendan, J. C., Pugh, C. M., Minter, R. M., Dunnington, G. L., & Kozar, R. A. (2008). The Changing Face of Surgical Education: Simulation as the New Paradigm. Journal of Surgical Research, 147(2), 189–193. http://doi.org/10.1016/j.jss.2008.02.014

Simon, H., & Chase, W. (1973). Skill in Chess. American Scientist, 61(4), 394–403.

Sir Alfred Cuschieri. (2003). Lest we forget the surgeon. Seminars in Laparoscopic Surgery, 10(3), 141–8. http://doi.org/10.1016/j.racsoc.2004.11.008

Spruit, E. N., Band, G. P. H., Hamming, J. F., & Ridderinkhof, K. R. (2014). Optimal training design for procedural motor skills: a review and application to laparoscopic surgery. Psychological Research, 78(6), 878–891. http://doi.org/10.1007/s00426-013-0525-5

Staats, B. R., KC, D. S., & Gino, F. (2018). Maintaining Beliefs in the Face of Negative News: The Moderating Role of Experience. Management Science, 64(2), 804–824. http://doi.org/10.1287/mnsc.2016.2640

Stratton, S. M., Liu, Y.-T., Hong, S. L., Mayer-Kress, G., & Newell, K. M. (2007). Snoddy (1926) Revisited: Time Scales of Motor Learning. Journal of Motor Behavior, 39(6), 503–515. http://doi.org/10.3200/jmbr.39.6.503-516

Sweller, J. (2008). Evolutionary bases of human cognitive architecture. Proceeding of the Fourth International Workshop on Computing Education Research - ICER '08, 1–2. http://doi.org/10.1145/1404520.1404521

Szasz, P., Bonrath, E. M., Louridas, M., Fecso, A. B., Howe, B., Fehr, A., … Grantcharov, T. P. (2016). Setting Performance Standards for Technical and Nontechnical Competence in General Surgery. Annals of Surgery, 266(1), 1. http://doi.org/10.1097/SLA.0000000000001931

Tofel-Grehl, C., & Feldon, D. F. (2013). Cognitive Task Analysis-Based Training: A Meta-Analysis of Studies. Journal of Cognitive Engineering and Decision Making, 7(3), 293–304. http://doi.org/10.1177/1555343412474821

Touchie, C., & Ten Cate, O. (2016). The promise, perils, problems and progress of competency-based medical education. Medical Education, 50(1), 93–100. http://doi.org/10.1111/medu.12839

Tyler, J. A., Clive, K. S., White, C. E., Beekley, A. C., & Blackbourne, L. H. (2010). Current US Military Operations and Implications for Military Surgical Training. Journal of the American College of Surgeons, 211(5), 658–662. http://doi.org/10.1016/j.jamcollsurg.2010.07.009

van der Vleuten, C. P. M. M., & Schuwirth, L. W. T. T. (2005). Assessing professional competence: from methods to programmes. Medical Education, 39(3), 309–317. http://doi.org/10.1111/j.1365-2929.2005.02094.x

van Merriënboer, J. J. G., Clark, R. E., & de Croock, M. B. M. (2002). Blueprints for complex learning: The 4C/ID-model. Educational Technology Research and Development, 50(2), 39–61. http://doi.org/10.1007/BF02504993

Vedula, S. S., Ishii, M., & Hager, G. D. (2017). Objective Assessment of Surgical Technical Skill and Competency in the Operating Room. Annual Review of Biomedical Engineering, 19(1), 301–325. http://doi.org/10.1146/annurev-bioeng-071516-044435

Weinbergger, S. E., Duffey, F. D., & Cassel, C. K. (2005). “Practice Makes Perfect” . . . Or Does It? Annals of Internal Medicine, 142, 302–303.

Wiegmann, D. A., ElBardissi, A. W., Dearani, J. A., Daly, R. C., & Sundt, T. M. (2007). Disruptions in surgical flow and their relationship to surgical errors: An exploratory investigation. Surgery, 142(5), 658–665. http://doi.org/10.1016/j.surg.2007.07.034

Willis, R. E., Gomez, P. P., Ivatury, S. J., Mitra, H. S., & Van Sickle, K. R. (2014). Virtual reality simulators: Valuable surgical skills trainers or video games? Journal of Surgical Education, 71(3), 426–433. http://doi.org/10.1016/j.jsurg.2013.11.003

Yule, S., Flin, R., Paterson-Brown, S., & Maran, N. (2006). Non-technical skills for surgeons in the operating room: A review of the literature. Surgery, 139(2), 140–149. http://doi.org/10.1016/j.surg.2005.06.017

Yule, S., & Paterson-Brown, S. (2018). Surgeons ’ Non-technical Skills. Surg Clin N Am, 92(2012), 37–50. http://doi.org/10.1016/j.suc.2011.11.004

2. Marker-less hand motion kinematics of simulated surgical tasks for quantifying surgeon experience

2.0 Manuscript Information

This manuscript will be submitted to special issue of *Applied Ergonomics* on human factors/ergonomics in health and healthcare, in honor of the late Bentzi Karsh.

2.1 Abstract

This paper summarizes observed hand motion differences among 85 clinicians and students performing benchtop suturing tasks. Medical students (32), residents (41), attending surgeons (10), and retirees (2) were recorded on digital video while suturing on one of foam, pig feet, or porcine bowel tissues. Each clinician was classified as junior or senior, within their role. Utilizing custom software, the location of each of the participants hands were automatically recorded throughout each frame of the video, producing a rich spatiotemporal feature set for subsequent comparison across participants. Observed differences between experience levels within each setting are described, with emphasis given to trends associated with increasing tenure. Increasing clinician tenure was associated with conserved path length per cycle of the non-dominant hand on the foam simulation, significantly reducing from early medical students (mean = 73.63 cm, sd = 33.21 cm) to senior residents (mean = 46.16 cm, sd = 14.03 cm, p = 0.015), and again between senior residents and senior attendings (mean = 30.84 cm, sd = 14.51 cm, p = 0.045). Attendings also accelerated less with their non-dominant hand (mean = 16.27 cm/s^2, sd = 81.12 cm/s^2, p = 0.002) than senior residents (mean = 24.84 cm/s^2, sd = 68.29 cm/s^2, p = 0.002), despite similar cycle rates. Medical students moved their dominant hands slower (mean = 4.39 cm/s, sd = 1.73 cm/s, p = 0.033) than senior residents (mean = 6.53 cm/s, sd = 2.52 cm/s) while tying. These results suggest that increases at early stages of training are gained by

improved dominant hand function, while increases in later stages are characterized by efficiently distributing work between hands.

2.2 Background

Stefanidis and colleagues (2015) describe a simulation-based training mantra for the 21st century as a transition from: “see one, do one, teach one” to “see one, simulate many deliberately, do one.” As part of this effort, and commensurate with burgeoning “surgical data science” (Maier-Hein et al., 2017), video review of surgery (Xiao et al., 2007) has led to improvements in skills analysis (Berger, Gaster, & Lee, 2013), coaching (Greenberg et al., 2015; Y.-Y. Hu et al., 2012; Soucisse et al., 2017), and error detection (Law Forsyth et al., 2017). In reviewing surgical assessment-oriented technologies, Vedula et al. (2017) describe how many objective computer-aided technical skill evaluation (OCASE-T) technologies depend on robotics or laparoscopy.

Our approach to surgical analysis, in contrast, uses motion capture of surgeon hand movements to analyze differences commensurate with experience. Video recording of the surgeon’s hands while operating, needing no sensors or markers, offers many advantages in portability and scalability otherwise limited in successful robot-assisted surgical systems like ROVIMAS (Aggarwal et al., 2007), and ICSAD (Bann, Khan, & Darzi, 2003; Datta et al., 2002; Hayter et al., 2009). Such platforms consistently discriminate novices from experts (Aristotelis Dosis et al., 2005; Overby & Watson, 2014) using metrics like the number of hand movements (sometimes referred to as economy), and overall path length. Our group found similar results in previous work using marker-less motion tracking and observed that attending surgeons move their non-dominant hands more than residents while suturing, yet distribute workload between hands more evenly, and generally conserve motion while tying (Glarner et al., 2014). We have

subsequently identified differences in dominant hand motion by role and task in the operating room (Frasier et al., 2016), to predict expert-rated performance (Azari et al., 2017).

This study represents an application of marker-less hand motion tracking to generalize findings between novices and experts for a broader range of experience levels within controlled settings. Video records of bench-top simulations offer a repeatable environment in which to hone and refine hand-motion kinematics for common procedures. Previous studies examining surgical motion have simulated small bowel anastomoses and vein patch insertions (Datta, Bann, Mandalia, & Darzi, 2006; Watson, 2014), and interrupted suturing tasks with commonly accessible materials like foam, balloons, and tissue paper (D'Angelo, Rutherford, Ray, Laufer, et al., 2015; D'Angelo, Rutherford, Ray, Mason, & Pugh, 2015).

This goal of this study is to examine differences in hand motions commensurate with a continuous range of experience as surgeons perform simulated suturing tasks on different materials. Grounded in previous work, we hypothesize that more experienced participants will exhibit faster completion rates and higher economy of motion in each setting. We manually calculate cycle frequency to provide common comparison in path lengths between different techniques. Parallel work described by Azari (2018) explores automatic prediction of these cycle rates through common machine learning techniques to remove the need for human labeling in surgical video analysis.

2.3 Methods

2.3.1 Participants and Setting

This study compares the hand motion results of 85 participants completing common benchtop suturing tasks. Three tissue conditions were used: foam, porcine feet and porcine bowel. Participants performed both simple interrupted and running subcuticular suturing on foam and pig feet, and anastomoses on bowel tissue. Thirty-seven participants were recruited through

grand-rounds announcements, email lists and live announcements to complete simple interrupted suturing and running subcuticular suturing on foam. Twenty-six participants were recruited to allow video recording during training sessions while suturing on pig's feet, while twenty-two participants were recruited via email to complete porcine bowel anastomosis (Figure 4). Third year medical students, residents with three or less years, and attendings with less than six years in their current role were classified as "junior," within each respective role.

Each participant agreed to have their hand movements recorded on digital video while they performed the suturing tasks. The University of Wisconsin-Madison Institutional Review Board approved this study. The number of participants and their relative experience levels are listed in Table 2.

Cameras were positioned to observe the hands and working space of each participant, minimizing visibility of faces in each setting (Figure 5). Cameras collected orthogonal 2D planar video with 720 x 480 pixel resolution at 30 frames per second utilizing software developed by the Occupational Ergonomics and Biomechanics Laboratory at the University of Wisconsin-Madison to synchronously record multiple views.

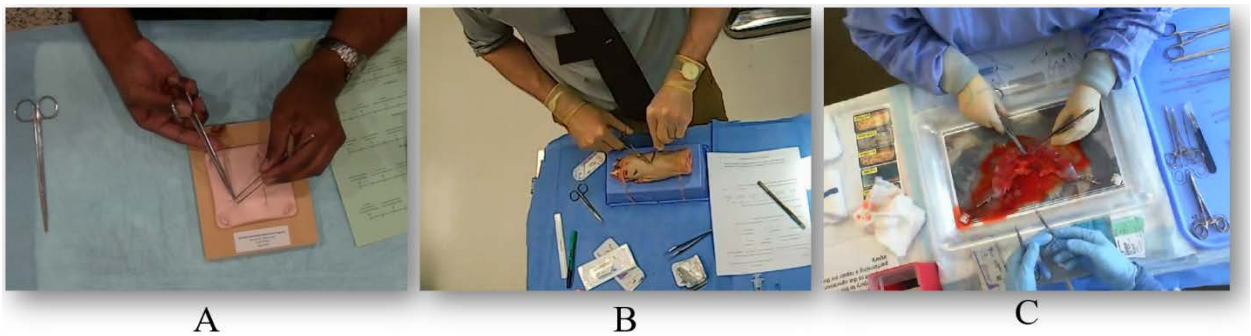


Figure 4: Example of camera view for training suturing tasks on foam (A), and porcine feet (B) and bowel (C).

Video recording began after reviewing and signing a consent agreement. Each video was calibrated to convert pixel measurements from the video into real-world (mm) units using the size of a known object (e.g. ruler or notecard) in view of the camera (Equation 1).

$$Calibration\ Coefficient = \frac{Millimeters\ per\ object\ dimension}{\#\ Pixels\ per\ object\ dimension} \quad (1)$$

Table 2: Number of participants, by role and tissue type.

Role	Year in Position	Foam Dressing (A)	Porcine Feet (B)	Porcine Bowel (C)	Participants
Medical Student (n = 32)	3	4	22	-	26
	4	6	-	-	6
Resident (n = 41)	1	2	2	3	7
	2	3	-	7	10
	3	3	2	10	15
	4	3	-	1	4
	5	4	-	1	5
Attending (n = 10)	< 6	4	-	-	4
	≥ 6	6	-	-	6
Retired (n = 2)	NA	2	-	-	2
Total		37	26	22	85

Extraneous footage (e.g. setting up, tearing down, time between stations) was trimmed from the video before motion tracking and labeling. However, there were occasional periods where participants would ask questions, discuss technique, remove their hands from the field or otherwise pause for long periods. These periods were manually identified and excluded from subsequent analysis.

2.3.2 Motion Tracking

A region of interest (ROI) in the video was defined over each of the participants' hands including the distal ends of two metacarpal bones. Utilizing custom software written in C# and employing the OpenCVSharp libraries (Chen, Hu, & Radwin, 2014), we were able to record and save the two-dimensional position of ROI's for both hands throughout the experiment. To

operate this software, an analyst would define the position and size of the ROI, initiate the algorithm, and provide manual corrections if the hands move off screen, or are otherwise occluded. The changing position of the ROI produced a unique position of each hand every 1/30th of a second, enabling speed, acceleration, and various other kinematic features to comprise a vector of attributes for each participant and task combination.

Figure 5: Video collection stations for two participants suturing on foam (left) and four participants suturing on pig feet (right).

2.3.3 Feature Extraction

From the two-dimensional position data for each frame it is possible to quantify instantaneous displacement, speed, and acceleration of both hands for each frame both instantaneously and over the course of a video Figure 6. Additional measures including jerk (Hogan & Sternad, 2009) and spatiotemporal curvature (Rao, Yilmaz, & Shah, 2002) are drawing increasing interest from research aiming to assess motion quality (Ghasemloonia et al., 2017). Jerk is the third derivative of position with respect to time and generally represents how smooth a motion is, while the spatiotemporal curvature function is a measure of direction change

based on multiple derivatives of the position signal and is used to indicate the number of discrete movements.

2.3.4 Cycle Analysis

After trimming extraneous portion of video, the remaining footage was screened in Multimedia Video Task Analysis (MVTA). MVTA is a software specially developed at the University of Wisconsin-Madison, (Yen & Radwin, 2007) to mark and save cycle starting and ending times for any task. Each frame of video was manually labeled in MVTA to provide ground truth for comparison across tasks and participants. A total of 10 states were identified, to provide sufficient resolution for state-prediction models (discussed in later chapters). These included suturing, tying (instrument, or one or two-handed), cutting, reach, maintaining tension, tissue manipulation, needle loading (or unloading) and extraneous/unrelated. The extraneous state comprised all periods of participant interaction, paperwork, significant pauses to ask questions or reposition equipment.

Each state included a series of motions, sometimes called “surgemes” (Lin, 2010). By convention, the surgical states identified here are commonly described as “maneuvers” within a broader series of “tasks” (e.g. closing an incision) and “procedures” (e.g. cholecystectomy) (Vedula, Malpani, Tao, et al., 2016). Progress in automatically predicting the arrival and transitions between these states are discussed by Azari (2018). The current paper uses labeled task breakdowns to segment the motion record and exclude unrelated activity.

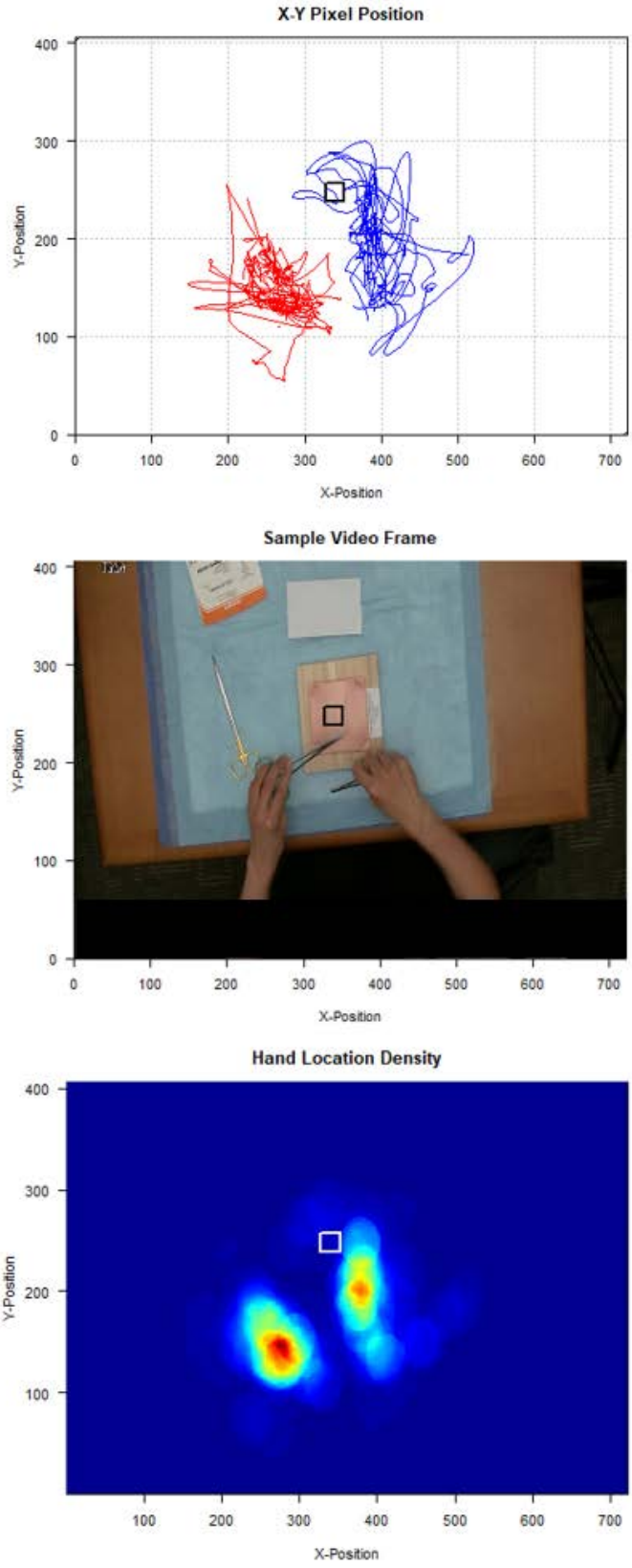


Figure 6: Example of data abstraction in X-Y pixel locations over time (top) and density of hand position over time (bottom) derived from video of hand motion (center).

2.3.5 Data Analysis

Following Shapiro-Wilks tests for normality, one-way ANOVA (analysis of variance) tests were performed to examine the impact of experience on hand motion within each one of the experimental settings. Interactions across settings were not within the scope of this paper. The amount of experience was the independent variable, with a kinematic feature as the dependent variable. Where significant differences were found, a Tukey Honestly Significant Difference (HSD) test and confidence intervals are used to describe the effect sizes between each of the respective groups. F values are reported for the grouped comparison. Kruskal-Wallis tests were used under non-normal or heteroscedastic (unequal variance) conditions to test for overall differences, with non-parametric pair-wise Wilcox tests examining differences between groups. We accounted for multiple comparisons using the Benjamini and Hochberg p-value correction. These tests are oriented to establish content evidence in accordance with Kane's framework (Cook et al., 2014) and compare performance with greater granularity than traditionally seen with binary experienced and novice distinctions. Features common to increasing tenure across different experimental settings are reported. For the purposes of this paper, ground-truth cycle frequency is used to compare experience categories, as well as to standardize comparisons of path length over the course of different length tasks. Future work examines automatic frequency calculation.

2.4 Results

The kinematic features exhibiting distinct trends are represented in Table 3. These include increasing cycle frequency (CF, Figure 7), decreasing path length per cycle (PLC, Figure 8), and changes in mean speed for dominant (D) and non-dominant (ND) hands. ND hand acceleration and standard deviation of speed also exhibits several differences.

Table 3: Features exhibiting significant differences by experience level. P = pigs feet; B = porcine bowel; F = foam dressing; A = combined (all) settings; M = medical student; JS = Junior medical student; SS = senior medical student; R = resident; A = attending; JR = junior resident; SR = senior resident; JA = junior attending; SA = senior attending. Pk = peak, Accl. = acceleration.

Feature	Hand	Settings	Task	*F* value	Significant Comparisons	*p* value
Cycle frequency (Hz)	Both	FD	All	4.52	M – SR	0.01
Cycle frequency (Hz)	Both	FD	Tying	10.24	A – JR	0.031
Cycle frequency (Hz)	Both	FD	Tying	9.66	JS – All	< 0.01
Cycle frequency (Hz)	Both	FD	Tying	9.66	JR – SA	0.013
Cycle frequency (Hz)	Both	FD	Suturing	6.45	SS – SR	0.001
Cycle frequency (Hz)	Both	FD	Suturing	9.55	JR – A	0.03
Cycle frequency (Hz)	Both	PF	All	34.40	JS – All	< 0.01
Path length per cycle (PLC)	ND	FD	All	6.88	SR – A	0.045
Path length per cycle (PLC)	ND	FD	All	5.72	SR – JS	0.015
Path length per cycle (PLC)	ND	FD	Active	5.58	SR – A	0.049
Path length per cycle (PLC)	ND	PF	All	3.04	JS – R	< 0.01
Path length per cycle (PLC)	ND	BA	Suturing	6.52	SR – JR	0.019
Path length per cycle (PLC)	D	FD	All	6.46	M – SR	< 0.01
Path length per cycle (PLC)	D	PF	All	3.30	JS – R	< 0.01
Path length per cycle (PLC)	D	BA	Suturing	5.40	SR – JR	0.031
Median Speed (mm/s)	ND	PF	Tying	23.06	JR – SR	0.019
Median Speed (mm/s)	ND	PF	Tying	23.06	M – SR	< 0.01
Median Speed (mm/s)	ND	FD	Tying	4.24	M – SR	0.033
Median Speed (mm/s)	ND	FD	Tying	4.24	SR – A	0.035
Median Speed (mm/s)	D	PF	All	6.48	M – R	< 0.015
Maximum Accl. (mm/s^2)	ND	FD	All	7.16	SR – A	0.001
Smooth Accl. Pk. Rate (Hz)	D	FD	All	1.96	M – SR	0.01
Smooth Accl. Pk. Rate (Hz)	D	PF	Tying	15.79	JS – SR	< 0.01
Smooth Accl. Pk. Rate (Hz)	D	PF	Tying	15.79	JR – SR	0.019

Cycle Frequency

Mean cycle frequency for all tasks on foam increased across student (mean = 0.12 Hz, sd = 0.05 Hz) and resident populations (mean = 0.17, sd = 0.06, $p = 0.03$), but plateaued between senior residents and attending roles (mean = 0.18, sd = 0.06). While tying on foam, junior medical students (mean = 0.12 Hz, sd = 0.05 Hz) were significantly slower ($p < 0.01$) than all other groups. Senior medical students (mean = 0.21 Hz, sd = 0.05 Hz) tied at similar rates to junior residents (mean = 0.22 Hz, sd = 0.07 Hz), and senior residents (mean = 0.27 Hz, sd = 0.08

Hz) tied similarly to junior attendings (mean = 0.25 Hz, sd = 0.07 Hz). Senior attendings (mean = 0.31 Hz, sd = 0.06 Hz), however, tied significantly faster ($p = 0.002$) than junior residents (Figure 7).

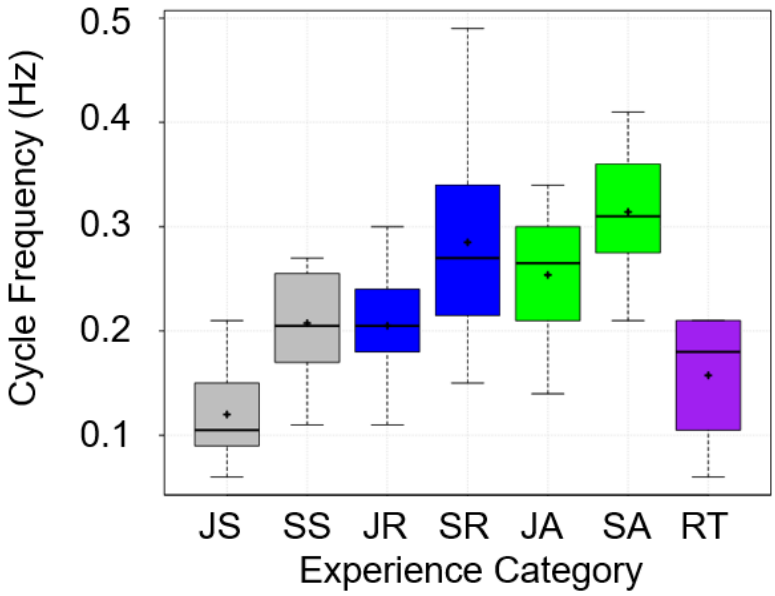


Figure 7: Cycle frequency for tying tasks on foam by experience level. Means are marked by "+". (J = Junior medical student (8); SS = senior medical student (12); JR = junior resident (10); SR = senior resident (20); JA = junior attending (8); SA = senior attending (12); RT = retired (4)).

Path Length

Differences in observed path length per cycle (PLC) were the most pronounced for non-dominant (ND) hand use while sewing on foam (Figure 8). Senior attendings (mean = 30.84 cm, sd = 14.51 cm) exhibited a slightly significant ($p = 0.045$) reduction in path length per cycle compared to senior residents (mean = 46.16 cm, sd = 14.03 cm), who exhibited significantly less PLC-ND than junior medical students (mean = 73.63 cm, sd = 33.21 cm, $p = 0.015$).

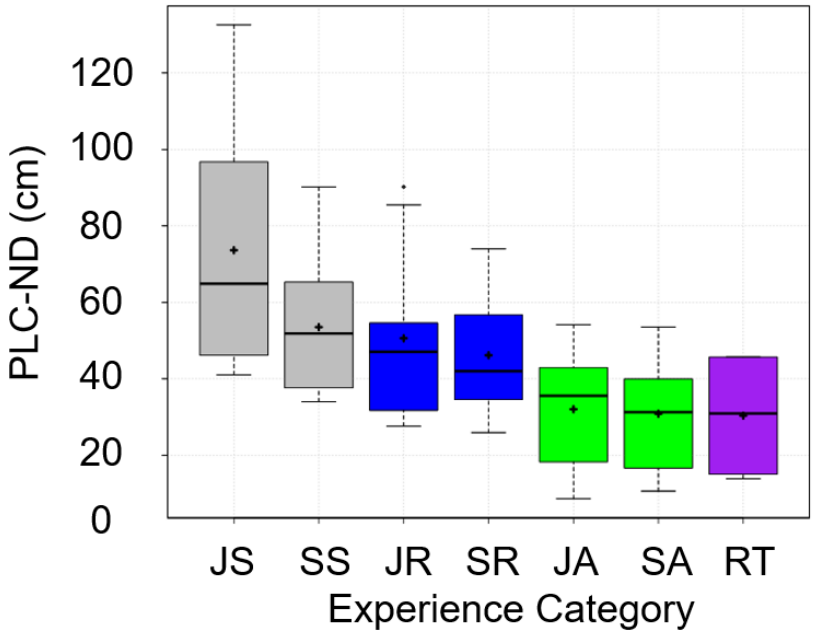


Figure 8: Path length per cycle of non-dominant hand (PLC-ND) use for simulated foam cases. Means are marked by "+". (J = Junior medical student (8); SS = senior medical student (12); JR = junior resident (10); SR = senior resident (20); JA = junior attending (8); SA = senior attending (12); RT = retired (4)).

Attending clinician PLC-ND excluding transitional periods for needle reloading or tissue repositioning (mean = 31.33 cm, sd = 15.29 cm) was also slightly lower ($p = 0.049$) than for senior residents (mean = 47.53 cm, sd = 15.67 cm). PLC-ND for these periods was also monotonic decreasing across experience categories, with the mean of retired samples (mean = 32.85 cm, sd = 18.19 cm, n = 4) close to junior attendings (mean = 32.11 cm, sd = 17.06 cm).

On pig feet, medical students (mean = 98.07 cm, sd = 79.28) had higher PLC-ND ($p < 0.01$) than both junior (mean = 46.7 cm, sd = 18.55 cm) and senior residents (mean = 50.66 cm, sd = 16.15 cm), although the difference within the resident population was insignificant. For bowel tissue, reduced PLC-ND was significant for active periods of suturing ($p < 0.02$) for senior residents (mean = 205.71 cm, sd = 82.88 cm) compared to junior residents (mean = 316.47 cm, sd = 120.15 cm), but not for tying, or for the overall task.

PLC of the dominant hand (PLC-D) while sewing on foam decreased across experience categories, with medical students exhibiting the highest path length (mean = 69.12 cm, sd = 28.67 cm), through residents (mean = 43.04 cm, sd = 15.80 cm), to attendings (mean = 40.32 cm, sd = 18.2 cm) and retirees (mean = 38.90 cm, sd = 19.15 cm). Senior resident PLC-D was significantly less than medical student PLC-D ($p < 0.01$), but the observed decreases from junior to senior resident, and from senior resident to attending were not significant.

Junior medical students (mean = 86.37 cm, sd = 59.57 cm) PLC-D on pig feet was similarly less than the combined resident population (mean = 47.74 cm, sd = 13.26 cm, $p < 0.01$). Senior resident PLC of the dominant hand (PLC-D) (mean = 264.55 cm, sd = 154.16 cm) while sewing on bowel, was significantly lower ($p = 0.031$) than junior resident PLC-D (mean = 391.33 cm, sd = 83.79 cm), despite greater standard deviation within the senior resident population.

Speed

Median speed of the non-dominant hand of junior residents (mean = 6.66 cm/s, sd = 2.26 cm/s) was significantly lower ($p = 0.019$) than median speed of senior residents (mean = 9.29 cm/s, sd = 3.06 cm/s) while tying on pig feet, but the differences for senior residents (mean = 4.3 cm/s, sd = 1.6 cm/s) and junior residents (mean = 4.21 cm/s, sd = 1.59 cm/s) tying on the more friable bowel material were not significant. While tying on foam, meanwhile, median speed of the non-dominant hand significantly increased ($p = 0.033$) from medical students (mean = 4.39 cm/s, sd = 1.73 cm/s) to senior residents (mean = 6.53 cm/s, sd = 2.52 cm/s), but attending median speed (mean = 4.41 cm/s, sd = 2.41 cm/s) resembled that of medical students and was significantly lower than the senior residents ($p = 0.035$, Figure 9).

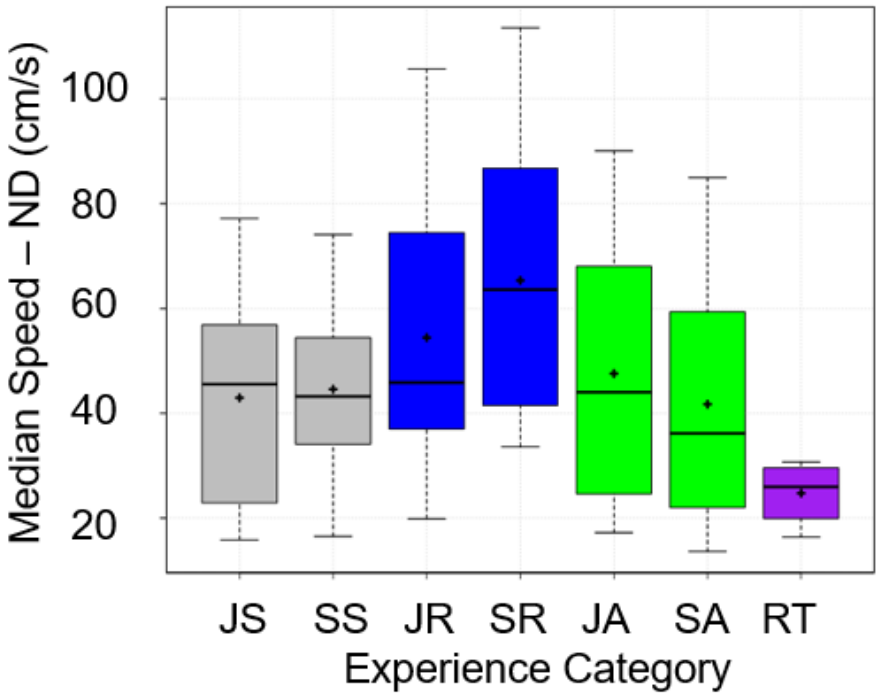


Figure 9: Median speed of non-dominant hand (ND) use for simulated foam cases. Means are marked by "+". (J = Junior medical student (8); SS = senior medical student (12); JR = junior resident (10); SR = senior resident (20); JA = junior attending (8); SA = senior attending (12); RT = retired (4)).

Resident sewing on pig feet exhibited greater average dominant (D) hand speed (mean = 7.67 cm/s, 8.25 cm/s, sd = 2.03 cm/s, 1.60 cm/s) than medical students (mean = 5.97 cm/s, sd = 1.53 cm/s, $p < 0.015$), but differences within the resident population were insignificant.

Acceleration

Maximum (90^{th} percentile) acceleration (ND) while sewing on foam increased from medical students (mean = 18.90 cm/s^2, sd = 60.07 cm/s^2) through junior (mean = 20.01 cm/s^2, sd = 69.30 cm/s^2) and senior (mean = 24.84 cm/s^2, sd = 68.29 cm/s^2) residents, but significantly declined between senior residents and attendings (mean = 16.27 cm/s^2, sd = 81.12 cm/s^2, p = 0.002), and continued to decline in retirement (mean = 98.92 cm/s^2, sd = 21.38 cm/s^2).

Like cycle frequency, the median Butterworth smoothed acceleration peak rate (D) for sewing on foam increased across early experience categories (Figure 10). Medical student peak

rates (mean = 0.18 Hz, sd = 0.17 Hz) were significantly slower than senior residents (mean = 0.26 Hz, sd = 0.07 Hz, $p = 0.01$). The increase observed at the entry levels, however, tapered off for junior (median = 0.28 Hz, sd = 0.11 Hz) and senior attendings (mean = 0.25 Hz, sd = 0.12 Hz) and again reduced for retirees (mean = 0.22 Hz, sd = 0.12 Hz).

While tying on pig feet, a similar trend was observed as medical students (mean = 1.09 Hz, sd = 0.61 Hz) became junior (mean = 1.52 Hz, sd = 0.68 Hz) and senior residents (mean = 2.41 Hz, sd = 0.80 Hz). There were significant differences between the senior residents and medical students ($p < 0.01$), and within the resident population ($p = 0.019$). Since no attendings were involved in tying on pig's feet, however, it is not possible to examine the subsequent trend to confirm a later reduction in peak acceleration arrival rates for this setting.

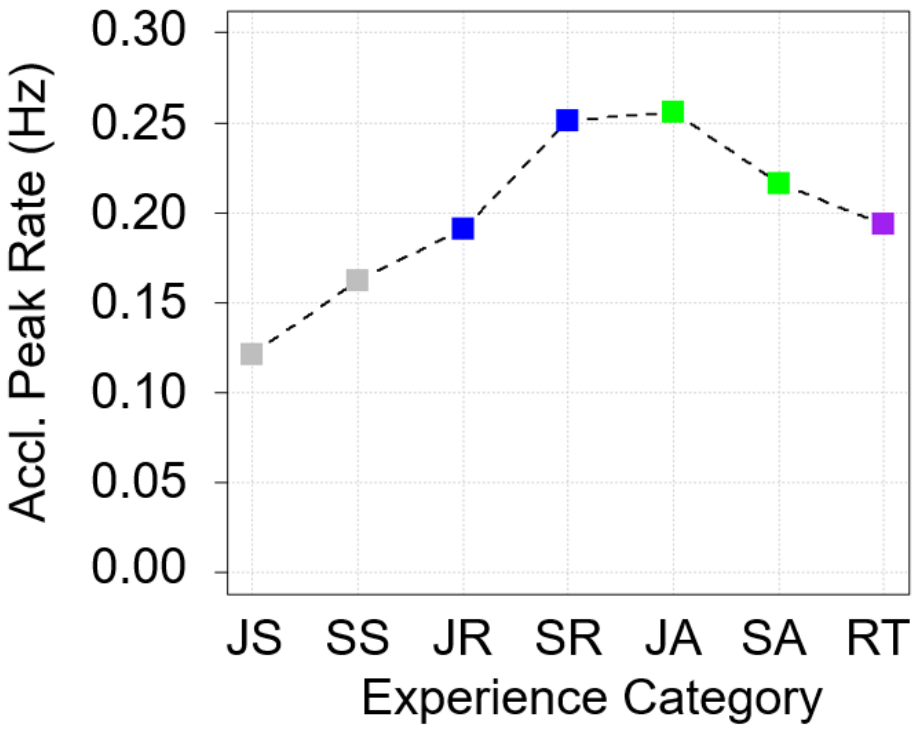


Figure 10: Median smoothed acceleration (accl.) peak arrival rate (Hz) for dominant hands by experience category. (J = Junior medical student (63); SS = senior medical student (12); JR = junior resident (28); SR = senior resident (43); JA = junior attending (8); SA = senior attending (12); RT = retired (4)).

2.5 Discussion

This study identifies features commensurate with increasing surgical tenure. Medical students, residents, and attendings were each classified as junior or senior, depending on their time in position. Retirees were treated as one group and, due to the limited sample size in that population, excluded from statistical testing. Members of each tier completed common suturing and tying maneuvers, and all groups sutured on foam. Only medical students and residents sutured on pig feet and porcine bowel. The range of experiences in the study provided greater resolution than the traditional distinction between novices and experts. In comparing trends of features across all experience levels, it is common that comparisons are significant only for non-adjacent categories.

Medical students and residents often exhibited differences in speed and acceleration for dominant hand use, while attendings and residents exhibited differences in their non-dominant hand for path length, speed, and acceleration. For suturing on foam, non-dominant hand speed increased as medical students became residents, and then reduced as residents became attendings, despite faster completion times. This may reflect a transition from learning to perform the task (medical student) and completing the task quickly (residents) to conserving energy and motion in performing a short familiar exercise (attendings).

The observed differences in cycle frequency, path length per cycle, median speeds, and acceleration suggest a pattern of increasing efficiency of movement along with tenure. Attendings exhibited greater cycle frequency, and less path length per cycle (PLC) for both dominant and non-dominant hands. Despite increased frequency, attendings also exhibited reduced speed of the non-dominant hand compared to residents while tying on foam. The smoothed acceleration signal for the dominant hand differentiated between medical students and residents (on foam), and within the resident population (on bowel).

Path length per cycle (PLC) of the surgeon's non-dominant (ND) hand showed a monotonic reduction across each experience category, with significant differences between attendings and senior residents, and between senior residents and medical students. Median speeds and maximum acceleration generally increased as medical students became residents but decreased across junior and senior attending roles. The peak arrival rate in the smoothed acceleration signal closely resembled the cycle frequency trend, with significant differences between medical students and residents, and diminishing increases once surgeons entered the attending role. Future study will examine how well the peak arrival rate in the acceleration signal can serve as an effective proxy for cycle frequency.

The limited number of retired surgeons in this study limits inference on skill decay beyond attending roles and into retirement. Deliberate practice may stave off effects of aging in psychomotor performance (Ericsson, 2004), but the rate of decay is thought to be independent of individual aptitude (Schendel et al., 1974). Future work would benefit from examining how the features identified in this study change following retirement or change of professional role.

Previous studies found that attendings, in general, use their non-dominant hands more than residents, yet reduce movement and conserve path length when appropriate (Glarner et al., 2014). In a similar light, we observed a decreasing path length per cycle in attendings non-dominant hands, accompanied by a reduction in maximum acceleration as residents became attendings. Increasing acceleration and hand speed prior to attending role were seen for foam and pig feet, but not for suturing on friable bowel material. The difficulty of this material may have prompted senior residents to slow down and spend greater time planning than they did for the more familiar materials. This may also suggest a greater amount of comfort and familiarity with the task surroundings and tool placement than their less experienced counterparts.

Davids et al., (2008) describes psychomotor performance in context of achieving stable "states" within a dynamic landscape of options. The increased acceleration within the resident population compared to attendings, could be an attempt to expedite time spent in familiar territory (practiced motions); a mirror image of Mouton's popular idea of "slowing down" to remain attentive (Moulton, Regehr, Lingard, Merritt, & MacRae, 2010). The increased pace during familiar portions may be rewarded by additional opportunity to "slow down" later and facilitate planning, decision making or error recovery, as referenced within Madani's surgical performance domain framework.

It is also difficult to rule out the possibility that clinicians of different standing intentionally sped up, slowed down, or altered their technique due to being aware of the video recording. Residents may have felt compelled to move noticeably faster than the medical students they train, and attendings with greater equanimity overall.

While we collected surgical motion across three settings, this study may not have targeted sufficiently difficult tasks to discriminate between junior and senior residents attendings in all cases. Significant differences observed within residents tying on friable bowel, but not on pig feet, could be an example of this principle. There may be similar features between attendings which are not as readily observable in the current scheme. More difficult tasks may demand proficiency in different domains (e.g. advanced cognitive skills) which may not be detected by our motion tracking algorithms in non-stressful operating situations. In other words, the "fundamentally different" approach and knowledge structure that an expert brings to the task (Prietula et al., 2000; Silber & Foshay, 2009), may not be observable in these kinds of benchtop tasks, or within the scope of motion tracking for randomly sampled video segments.

The kind of motion analysis applied in this study also does not account for successfully completed procedures, and rests on the assumption that all participants completed the task.

Future regression analyses or deep learning algorithms may take advantage of the observable hand motion, however, to predict more about relative performance and contextual state. Automatic routines to predict performance and identify periods of suturing, tying, or transitional activity from raw video are explored in accompanying work (Azari, 2018) currently under review.

Creating a motion record for each participant and ensuring that all periods of extraneous activity were accounted for, including out of frame motion, proved to be the most time-consuming portion of this study. Rapid changes in viewable hand size and shape, for instance, caused the motion tracking algorithm to lose track, and required manual intervention. Despite these current challenges, computer vision capabilities will continue to improve and reduce the burden to apply motion tracking. We have recently enhanced our ability to supervise the ROI throughout a video with a new interface design. We have also implemented simultaneous multiple ROI tracking for one video. These advantages will not correct for out of frame motion irregular behavior, or changing hand shape, but they will decrease the number of passes needed to create a motion record and reduce the burden of checking and controlling for extraneous activity. The software is also designed for modularity, in that pixel information can be passed on-demand to any selected algorithm, promoting further algorithm refinement.

2.6 Conclusion

This study explored hand motion features associated with increasing surgeon experience. Participants from six experienced categories completed common suturing tasks on three kinds of tissue. Increasing tenure was associated with greater cycle frequency, decreased path length per cycle for both hands, increased speed and acceleration as medical students became residents, but reducing speed and acceleration for attending surgeons. The peak arrival rate in the smoothed acceleration signal may be a proxy for cycle frequency and should be explored further in future

work. Taken as an ensemble, the features identified in this study describe how marker-less motion tracking can quantify “surgical dexterity” for simulated benchtop tasks in various settings and for a range of experience levels.

2.7 References

Aggarwal, R., Grantcharov, T., Moorthy, K., Milland, T., Papasavas, P., Dosis, A., … Darzi, A. (2007). An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Annals of Surgery, 245(6), 992–999. http://doi.org/10.1097/01.sla.0000262780.17950.e5

Azari, D. P. (2018). Quantifying Surgical Skill. University of Wisconsin-Madison.

Azari, D. P., Frasier, L. L., Quamme, S. R. P., Greenberg, C. C., Pugh, C. M., Greenberg, J. A., & Radwin, R. G. (2017). Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. Annals of Surgery, XX(Xx), 1. http://doi.org/10.1097/SLA.0000000000002478

Bann, S. D., Khan, M. S., & Darzi, A. W. (2003). Measurement of surgical dexterity using motion analysis of simple bench tasks. World Journal of Surgery, 27(4), 390–394. http://doi.org/10.1007/s00268-002-6769-7

Berger, A. J., Gaster, R. S., & Lee, G. K. (2013). Development of an Affordable System for Personalized Video-Documented Surgical Skill Analysis for Surgical Residency Training. Annals of Plastic Surgery, 70(4), 442–446. http://doi.org/10.1097/SAP.0b013e31827e513c

Chen, C.-H., Hu, Y. H., & Radwin, R. G. (2014). A motion tracking system for hand activity assessment. In 2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP) (pp. 320–324). IEEE. http://doi.org/10.1109/ChinaSIP.2014.6889256

Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Advances in Health Sciences Education, 19(2), 233–250. http://doi.org/10.1007/s10459-013-9458-4

D’Angelo, A. L. D., Rutherford, D. N., Ray, R. D., Laufer, S., Kwan, C., Cohen, E. R., … Pugh, C. M. (2015). Idle time: An underdeveloped performance metric for assessing surgical skill. American Journal of Surgery, 209(4), 645–651. http://doi.org/10.1016/j.amjsurg.2014.12.013

D’Angelo, A. L. D., Rutherford, D. N., Ray, R. D., Mason, A., & Pugh, C. M. (2015). Operative skill: Quantifying surgeon’s response to tissue properties. Journal of Surgical Research, 198(2), 294–298. http://doi.org/10.1016/j.jss.2015.04.078

Datta, V., Bann, S., Mandalia, M., & Darzi, A. (2006). The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. American Journal of Surgery, 192(3), 372–378. http://doi.org/10.1016/j.amjsurg.2006.06.001

Datta, V., Chang, A., Mackay, S., & Darzi, A. (2002). The relationship between motion analysis and surgical technical assessments. The American Journal of Surgery, 184(1), 70–73. http://doi.org/10.1016/S0002-9610(02)00891-7

Davids, K., Button, C., & Bennett, S. (2008). Dynamics of Skill Acquition - A Constraints-Led Approach. Champaign, IL, USA: Human Kinetics.

Dosis, A., IK, S., AA, F., M, M., K, Y., A, D., & Darzi, A. (2005). Synchronized Video and Motion Analysis for the Assessment of Procedures in the Operating Theater. Archives of Surgery, 140(3), 293. http://doi.org/10.1001/archsurg.140.3.293

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Academic Medicine : Journal of the Association of American Medical Colleges, 79(10 Suppl), S70–S81. http://doi.org/10.1097/00001888-200410001-00022

Frasier, L. L., Azari, D. P., Ma, Y., Quamme, S. R. P., Radwin, R. G., Pugh, C. M., … Greenberg, C. C. (2016). A marker-less technique for measuring kinematics in the operating room. Surgery. http://doi.org/10.1016/j.surg.2016.05.004

Ghasemloonia, A., Maddahi, Y., Zareinia, K., Lama, S., Dort, J. C., & Sutherland, G. R. (2017). Surgical Skill Assessment Using Motion Quality and Smoothness. Journal of Surgical Education, 74(2), 295–305. http://doi.org/10.1016/j.jsurg.2016.10.006

Glarner, C. E., Hu, Y. Y., Chen, C. H., Radwin, R. G., Zhao, Q., Craven, M. W., … Greenberg, C. C. (2014). Quantifying technical skills during open operations using video-based motion analysis. Surgery (United States), 156(3), 729–734. http://doi.org/10.1016/j.surg.2014.04.054

Greenberg, C. C., Ghousseini, H. N., Pavuluri Quamme, S. R., Beasley, H. L., & Wiegmann, D. A. (2015). Surgical Coaching for Individual Performance Improvement. Annals of Surgery, 261(1), 32–34. http://doi.org/10.1097/SLA.0000000000000776

Hayter, M. A., Friedman, Z., Bould, M. D., Hanlon, J. G., Katznelson, R., Borges, B., & Naik, V. N. (2009). Validation of the Imperial College Surgical Assessment Device (ICSAD) for labour epidural placement. Canadian Journal of Anesthesia, 56(6), 419–426. http://doi.org/10.1007/s12630-009-9090-1

Hogan, N., & Sternad, D. (2009). Sensitivity of smoothness measures to movement duration, amplitude, and arrests. Journal of Motor Behavior, 41(6), 529–34. http://doi.org/10.3200/35-09-004-RC

Hu, Y.-Y., Peyre, S. E., Arriaga, A. F., Osteen, R. T., Corso, K. A., Weiser, T. G., … Greenberg, C. C. (2012). Postgame Analysis: Using Video-Based Coaching for Continuous

Professional Development. Journal of the American College of Surgeons, 214(1), 115–124. http://doi.org/10.1016/j.jamcollsurg.2011.10.009

Law Forsyth, K., DiMarco, S. M., Jenewein, C. G., Ray, R. D., D'Angelo, A.-L. D., Cohen, E. R., … Pugh, C. M. (2017). Do errors and critical events relate to hernia repair outcomes? The American Journal of Surgery, 213(4), 652–655. http://doi.org/10.1016/j.amjsurg.2016.11.020

Lin, H. C. (2010). Structure in surgical motion. Johns Hopkins University.

Maier-Hein, L., Vedula, S., Speidel, S., Navab, N., Kikinis, R., Park, A., … Jannin, P. (2017). Surgical Data Science: Enabling Next-Generation Surgery. Retrieved from https://arxiv.org/abs/1701.06482

Moulton, C. A., Regehr, G., Lingard, L., Merritt, C., & MacRae, H. (2010). Slowing down to stay out of trouble in the operating room: remaining attentive in automaticity. Acad Med, 85(10), 1571–1577. http://doi.org/10.1097/ACM.0b013e3181f073dd

Overby, D. W., & Watson, R. A. (2014). Hand motion patterns of Fundamentals of Laparoscopic Surgery certified and noncertified surgeons. The American Journal of Surgery, 207(2), 226–230. http://doi.org/10.1016/j.amjsurg.2013.10.007

Prietula, M. J., Feltovich, P. J., & Marchak, F. (2000). Factors influencing analysis of complex cognitive tasks: a framework and example from industrial process control. Human Factors, 42(1), 56–74. http://doi.org/10.1518/001872000779656589

Rao, C., Yilmaz, A., & Shah, M. (2002). View-invariant representation and recognition of actions. International Journal of Computer Vision, 50(2), 203–226. http://doi.org/10.1023/A:1020350100748

Schendel, J. D., Shields, J. L., & Katz, M. G. (1974). Retention of motor skills: Review. Technical Paper, 50.

Silber, K. H., & Foshay, W. R. (2009). Handbook of Improving Performance in the Workplace, Instructional Design and Training Delivery (1st ed.). San Francisco: Pfeiffer.

Soucisse, M. L., Boulva, K., Sideris, L., Drolet, P., Morin, M., & Dubé, P. (2017). Video Coaching as an Efficient Teaching Method for Surgical Residents—A Randomized Controlled Trial. Journal of Surgical Education, 74, 365–371. http://doi.org/10.1016/j.jsurg.2016.09.002

Stefanidis, D., Sevdalis, N., Paige, J., Zevin, B., Aggarwal, R., Grantcharov, T., … Association for Surgical Education Simulation Committee. (2015). Simulation in surgery: what's needed next? Annals of Surgery, 261(5), 846–53. http://doi.org/10.1097/SLA.0000000000000826

Vedula, S. S., Ishii, M., & Hager, G. D. (2017). Objective Assessment of Surgical Technical Skill and Competency in the Operating Room. Annual Review of Biomedical Engineering, 19(1), 301–325. http://doi.org/10.1146/annurev-bioeng-071516-044435

Vedula, S. S., Malpani, A. O., Tao, L., Chen, G., Gao, Y., Poddar, P., … Chen, C. C. G. (2016). Analysis of the Structure of Surgical Activity for a Suturing and Knot-Tying Task. PLOS ONE, 11(3), e0149174. http://doi.org/10.1371/journal.pone.0149174

Watson, R. A. (2014). Use of a Machine Learning Algorithm to Classify Expertise: Analysis of Hand Motion Patterns During a Simulated Surgical Task. Academic Medicine : Journal of the Association of American Medical Colleges, 89(8), 1–5. http://doi.org/10.1097/ACM.0000000000000316

Xiao, Y., Schimpff, S., Mackenzie, C., Merrell, R., Entin, E., Voigt, R., & Jarrell, B. (2007). Video technology to advance safety in the operating room and perioperative environment. Surgical Innovation, 14(1), 52–61. http://doi.org/10.1177/1553350607299777

3. Using surgeon hand motions to predict surgical maneuvers

3.0 Manuscript Information

This manuscript will be submitted to the journal *Human Factors.*

3.1 Abstract

Automatic computer vision recognition of surgical maneuvers (e.g. suturing and tying) would expedite video review and support objective assessment. We recorded the hand movements of 37 clinicians performing simple and running subcuticular suturing benchtop simulations and applied three machine learning techniques (decision trees, random forests, and hidden markov models) to classify surgical maneuvers for every two seconds (60 frames) of video. Random forest predictions of surgical video into suturing, tying, and transition states correctly classified 74% of all video segments in a randomly selected test set. Hidden markov model adjustments improved the random forest predictions to 79% for simple interrupted suturing on a subset of randomly selected set of participants. These results enable automatic calculation of cycle frequency and path length per cycle – meaningful metrics in surgical skill and performance assessment.

3.2 Background

A surgical operation can be described as a series of procedures, tasks, maneuvers, and gestures. Vedula et al. (2016) provide a “hierarchical semantic decomposition of surgical activity” which defines a mutually exclusive set of terms to represent unique hand-tool movements of tasks within a procedure (e.g. appendectomy). A task necessary to complete a procedure (e.g. close incision) would include maneuvers like suturing (e.g. stitch) or tying (e.g. two-loop or one loop knot). Gestures would include several intermediate steps (Figure 11) within each maneuver, sometimes called “surgemes” (Lin, 2010), or “strokes” (Ahmidi et al., 2015).

Some studies further deconstruct “surgemes” into “dexemes” to facilitate highly granular segmentation (Despinoy et al., 2016).

Automatic classification of surgical procedures into similar terms through video would offer a more efficient after-action review; it would provide a “black box” to identify common motion patterns, and perhaps help identify errors or examples for future training. Automatic classification would support quantitative feedback during coaching sessions by comparing a learner’s motion trajectory throughout a procedure to a template “expert” trajectory – a common strategy for other psychomotor performance-based tasks like dance, soccer and tennis (Davids et al., 2008). Automatic state deconstruction further enables automatic quantification of cycle frequency. This removes the burdensome relationship in assessing overall path length – a common discriminator of skill (Aggarwal et al., 2007) – with the overall time of the procedure. Efforts to deconstruct surgical hand motion into gestures is part of a broader effort to develop an “ontological language of surgery” (Zappella, Béjar, Hager, & Vidal, 2013). Deconstructing surgical performance also represents a compelling problem for machine learning and image processing.

These kinds of gesture recognition and classification through computer vision are varied (Gavrila, 1999; Poppe, 2007; Wang, Hu, & Tan, 2003), and continue to grow. Motion chain-codes and recurrent neural networks, for example, are employed to recognize numbers traced by hand (Bhuyan, Ajay Kumar, MacDorman, & Iwahori, 2014), and maneuvers for robot-assisted suturing (Dipietro et al., 2016). Additional work by Reiley and colleagues (2008), in testing recognition of eleven “surgemes” performed while operating the da Vinci surgical system, classified more than 70% of tasks for participants of varying skill. The authors acknowledged the difficulty in addressing the variety of techniques participants exhibit while completing the same task. Ahmidi et al., (2017), similarly reported a 10% decrease in accuracy for user or participant

controlled cross validation in testing various state of the art methods for laparoscopic state prediction. Numerous approaches applying machine learning to surgical skill and state analysis, a portion of what is known more broadly as artificially intelligent medicine (Patel et al., 2009) and "surgical data science" (Maier-Hein et al., 2017), are identified in Table 4.

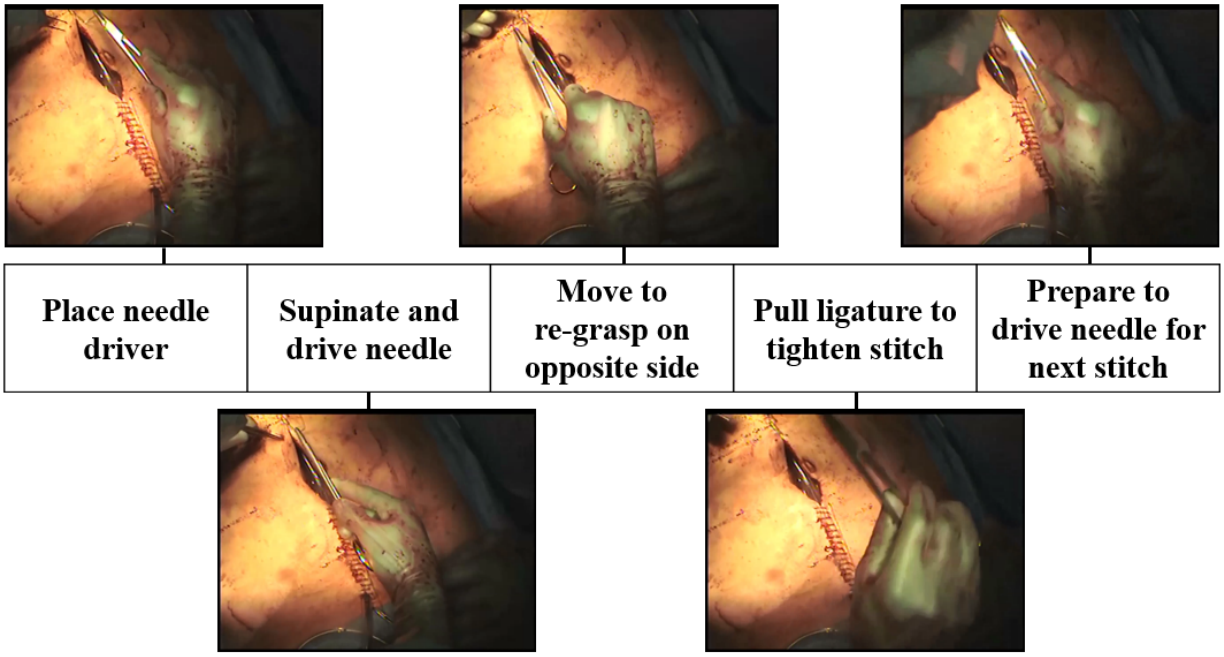


Figure 11: Observable "surgemes" for common tool-suturing technique while closing along the body wall. Completion represents one full cycle within the larger suturing maneuver.

Most instances of surgical gesture classification depend primarily on active sensor or robotic record tracking (Reiley et al., 2008). Indeed, there are increasing studies of the publicly accessible Gesture and Skill Assessment Working Set (JIGSAWS) (Gao et al., 2014) based on output by the da Vinci robot-assisted platform (Ahmidi et al., 2017; Lea, Hager, & Vidal, 2015; Lea, Vidal, & Hager, 2016). Ahmidi (2013), for example, used the da Vinci platform to predict three gestures ("grab", "pull", "rotate") for different performance levels with over 90% accuracy. Unsupervised approaches, including temporal clustering have demonstrated up to 88% classification accuracy of what the authors call "pseudo laparoscopic procedures" or "surgical phases" on training data (Zia, Zhang, Xiong, & Jarc, 2017) and 82% on testing data (Despinoy et

al., 2016). Automatic tool recognition through video to classify surgical state is also of growing interest (Bouget, Allan, Stoyanov, & Jannin, 2017).

Table 4: Pertinent machine learning algorithms for surgical state and skill analysis.

Supporting Literature	*Method*	*Common Uses*	*Rationale & Intended Use*
(Padoy et al., 2012)	Dynamic Time Warping (DTW)	Precursor to HMMs; can compare similarity of signals at different speeds	Symbolically group tying or suturing tasks together; similarity in motion signals may indicate different experience levels performing the same task
(Ahmidi et al., 2013)	Common String Models (CSMs)	Longest-string similarity comparisons	
(J. Rosen, Brown, Chang, Sinanan, & Hannaford, 2006)	k-means clustering	Unsupervised learning classification algorithm	Without relying on any ground-truth data, partition kinematics by experience levels and expert ratings
(Gaber, Zaslavsky, & Krishnaswamy, 2005)	k-Nearest Neighbors (k-NNs)	Supervised learning classification algorithm	Using ground-truth data, partition kinematics by experience levels and expert ratings
(Watson, 2014)	Support Vector Machines (SVMs)	Classification of test sequences into families	Use kinematic features and/or HMM representations to classify motion patterns by skill levels
(Fating & Ghotkar, 2014; Iivarinen & Visa, 1996)	Chain Code Histograms (CCHs)	Group together similar 2D representations	Associate shapes of surgeon hand motions by task, or experience levels within a task
(Uemura et al., 2014)	Detrended Fluctuation Analysis (DFA)	Assess adherence to repeated sequences	Examine similarity of kinematics for tasks of different individuals
	Unstable Periodic Orbit Analysis (UPOA)	Assess amount of stability in repeating sequences	
(Mackel, Rosen, & Pugh, 2007)	Hidden Markov Models (HMMs)	Speech recognition systems, state-based and sequential (hierarchical) pattern matching	Represent motions and transitions between motions as unique properties of a cohort (i.e. similarly rated and/or experienced)

While robotic and sensor-aided surgical gesture classification grows in complexity and accuracy, classification and assessment of hand motion during open procedures – those necessarily without sensors or robotic feedback – remain underdeveloped (Vedula et al., 2017).

The goal of this study is to classify surgical maneuvers from digital video of the hands with similar accuracy to existing studies of robot-assisted surgery. We explore the potential of decision trees, random forests and Hidden Markov Models (HMM) to appropriately distinguish between surgical maneuvers.

3.2.1 Motion Tracking

We have demonstrated how tracked hand motion in videos quantify kinematic properties of movements and exertions for specific tasks without special sensors or markers (Akkas et al., 2014; Azari et al., 2015; C. H. Chen et al., 2014), and explored novel visualization techniques to describe repetitive motion (Greene, Azari, Hu, & Radwin, 2017). Previous studies by our group applying this technology to surgical procedures have focused on testing the feasibility of marker-less video of motion analysis to isolate kinematic differences (i.e. displacement, speed, acceleration) (Glarner et al., 2014), predict performance (Azari et al., 2017), and identify meaningful differences between attendings and residents performing live surgery in the operating room (Frasier et al., 2016). These studies have established that marker-less video motion analysis of open procedures is feasible, and that it can identify differences in behavior between tasks and levels of experience.

3.2.2 Decision Trees and Random Forests

Decision trees are interpretable “white box” classification techniques that split data into categories based on simple rules, represented as an intelligible flowchart of if-then statements. Known as “greedy” algorithms, however, decision trees have also long been criticized for poorly balancing variance, bias, over-fitting and complexity (Barros, de Carvalho, & Freitas, 2015; Criminisi, 2011).

Random forests are “black box” ensembles of decision trees, intended to improve the accuracy and reliability of decision tree classification. A random forest is comprised of many

decision trees, each of which randomly select many subsets of features. Each tree "votes" for a classification outcome (Breiman, 2001). The final classification depends on the average of the forest, rather than on a single tree. Random forests provide additional advantage over other ensemble methods such as "bagging" (bootstrap aggregation) through selecting random subsets of features, and over single trees in general by partitioning the data to create multiple competing predictions – thereby lessening the "greediness" in early branches.

Despite these advantages, decision trees (ensemble or otherwise) do not retain temporal state information. The state prediction from a random forest for a period of video has no memory of the previous state, and no expectation for the following. A surgeon completing each task, on the other hand, is reasonably expected to progress through a procedure in a predictable pattern. As a result, this study applies HMMs, in combination with decision tree predictions, to improve surgical state classification.

3.2.3 Hidden Markov Models

HHMs are commonly used to analyze spoken (Rabiner, 1989), and sign languages (Starner & Pentland, 1995). Designed to examine "indirect evidence" associated with an underlying (hidden) state, HMMs may also predict the state of surgery by examining observable hand gestures. Such "gestemes" have successfully predicted human-machine interaction joint painting tasks (Hundtofte, Hager, & Okamura, 2002) and modeled sequences of hand movements during robotic surgery (Haro Bejar, Zappella, & Vidal, 2012; Tao & Elhamifar, 2012; Tao, Zappella, Hager, & Vidal, 2013; Zappella et al., 2013). HMMs have also demonstrated success in discriminating between novice and expert surgeons during laparoscopy (Rosen, Hannaford, Richards, & Sinanan, 2001).

However, designing HMMs is not without challenge. In 2007, Mackel, Rosen, & Pugh found that accuracy of HHMs was sensitive to the number of states chosen while predicting

experience from sensor data during an simulated pelvis exam. Representing a time series as a symbolic representation acceptable to model with an HMM is also an avenue of continuing research (Zucchini, MacDonald, & Langrock, 2016). This study explores the utility of HMMs to improve the classification of surgical video, without sensors or markers, during common benchtop suturing tasks.

HMMs are a probabilistic representation of a sequence of states. They consist of states, transitions between those states, and the probability of observing some feature associated with state (called an emission matrix). For classification of continuous time-series data, each record is first converted to a series of symbols before HMMs are trained in a supervised approach, where a log-likelihood trained model is applied to testing data (Zucchini et al., 2016). A surgeon's hand motion can be represented as a bivariate time series of step lengths and turning angles; and in the case of video, measured 30 times a second (Figure 12).

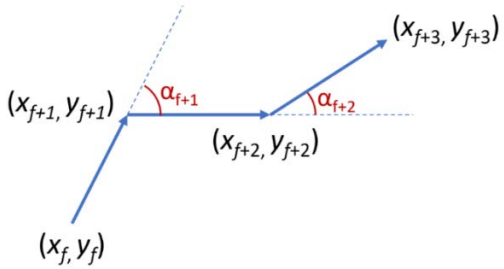


Figure 12: X and Y pixel locations at each frame (f), with deviations in angle (α) at every step.

The best HMM classification rates of robotically assisted surgery into three states is generally between 70 and 80 percent (Tao et al., 2013), with some periods of tying recognized as high as 93% when integrating sensor and video data into a hybrid classification approach (Zappella et al., 2013). Accuracy predicting surgemes during knot tying such as "both hands pull" have been classified correctly as high as 97% (Haro Bejar et al., 2012).

The goal of this study is to demonstrate that video motion capture can be used to predict surgical states with similar accuracy. We compare existing classification accuracy of robotic

three-state models to those derived only from digital video motion capture, quantifying both the number of time periods classified correctly and the number of cycles correctly identified in each video record.

3.3 Methods

3.3.1 Participants and Setting

We recruited clinicians of varying experience for this study to address high variability among individual and skill-levels observed in previous studies. Attending surgeons were recruited via email request and announcements during grand rounds, while resident and medical student participants were recruited through announcements during common surgical skills training sessions. A total of thirty-seven participants agreed to have their hand motions recorded on digital video while performing two common suturing tasks. Medical students (n=10), residents (n=15), attending surgeons (n=10), and retirees (n=2), completed three simple interrupted stiches, followed by a running subcuticular suture (approximately 5 cm in length). Residents who completed up to three post-graduate years (PGY) are classified as “junior residents” (n=5). Participation required 12-15 minutes, including review and signing of the consent agreement, completion of a demographic questionnaire, and video recording.

The two incisions (one for each task) were simulated by cuts (7.6 cm long) in an allevyn hydrocelluar foam dressing (10.2 cm x 10.2 cm), mounted to a wood block (15.2 cm x 15.2 cm) for stability. A small towel was placed under each dressing so that it would “pucker,” exposing the interior of the incision. Participants completed the simple interrupted suturing task on one incision, followed by the running subcuticular task on the other. Prior to the experiment, each participant completed a brief demographic survey detailing their surgical role and experience and reviewed the consent agreement. Participation and recruitment were approved by the Social and

Behavioral Health and Science Institutional Review Board at the University of Wisconsin-Madison.

3.3.2 Video Motion Tracking

Cameras were mounted overhead and positioned to observe only the participant's working area (see Figure 13). Faces were not visible. We used software that our group has developed for marker-less video processing single camera digital video to reliably track the motion trajectory of a selected region of interest over each video frame without the need for sensors or markers (C.-H. Chen, Hu, Yen, & Radwin, 2012; Chia-Hsiung Chen et al., 2015, 2014). Written in Matlab and C# with the open-source OpenCVSharp (.Net wrapper for the OpenCV) vision library, this software is based on a cross-correlation template matching algorithm which anticipates possible trajectories across the video (known as a sequential Bayesian estimation framework). Without any additional sensors or instruments to track hand motion, given a frame of video the software will save the spatiotemporal location of the ROI in that frame. The software allows us to identify an initial square region on an arbitrary moving object in a video clip such as the hand, called the region of interest (ROI), and track that object as it moves in the plane of view. The position of the ROI (seen in) is tracked across each frame of a video and stored within a vector in a unique data-frame. This approach enables mathematical abstraction of motion for subsequent pattern and feature analysis.

3.3.3 Surgical State Model

The tasks in this study were represented by a three-state model: (1) suturing, (2) tying, and (3) transition. A suturing maneuver began when a participant first touched tissue to drive a needle, continued while the participant pulled the ligature to the desired tension and ended when the ligature had reached its final position. Tying began at the first change in direction to initiate a knot and ends similarly when tension on the ligature is released to initiate the next gesture. Each

knot was marked as a unique tying cycle. The transitional state comprises switching from suturing to tying or vice-versa and includes reaching and cutting. The transition state may also encapsulate any other periods of extraneous activity (e.g. writing, adjusting the chair, filling out paper work, adjusting the simulation, or selecting a new suture or needle driver), where it occurs. The first labeled state began when the participant first touched the tissue to begin the task.

Figure 13: Video collection station (left) and region of interest (ROI, right) on participant's right hand, encompassing unique portion of hand.

The two-dimensional position record of both dominant and non-dominant hands (Figure 14), enable automatic calculation of numerous features including speed, acceleration, jerk, their fast-Fourier transforms (FFT), as well as the frequency and peak arrival rates of raw and butterworth-smoothed signals for both hands. Utilizing the position records relative to one another, the distance between the hands, relative distance from the simulation center, and hand angle are also computed, in addition to the speed and acceleration of these changes. In total, a feature set of 1213 predictor variables were computed for each video.

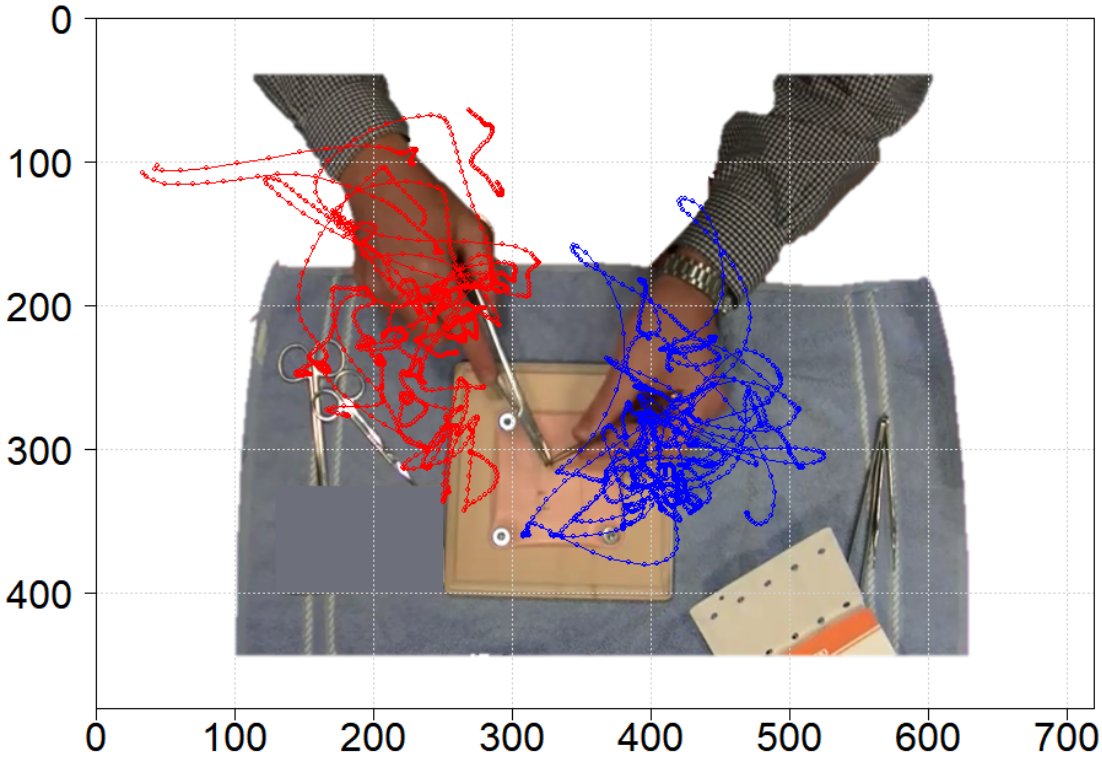


Figure 14: X and Y pixel location of the hands (background) over 30 seconds of a simple interrupted suturing task.

For a simple interrupted suturing task, a participant would generally repeat the following sequence: suture, tie, and transition (Figure 15). A transition typically included reaching, cutting, reloading the needle driver, and at times, a temporary pause in motion called “maintaining tension.” Reaching included periods of unloaded hand-movement, after which a new gesture began. Cutting included any time when the participant was holding scissors. Although this paper focuses on comparing accuracy of a three-state prediction to existing three-state robotic classifications at a maneuver level, future work may utilize these additional states test gesture-level classification methods. Running subcuticular suturing included several periods of suturing, followed by a period of tying and transition.

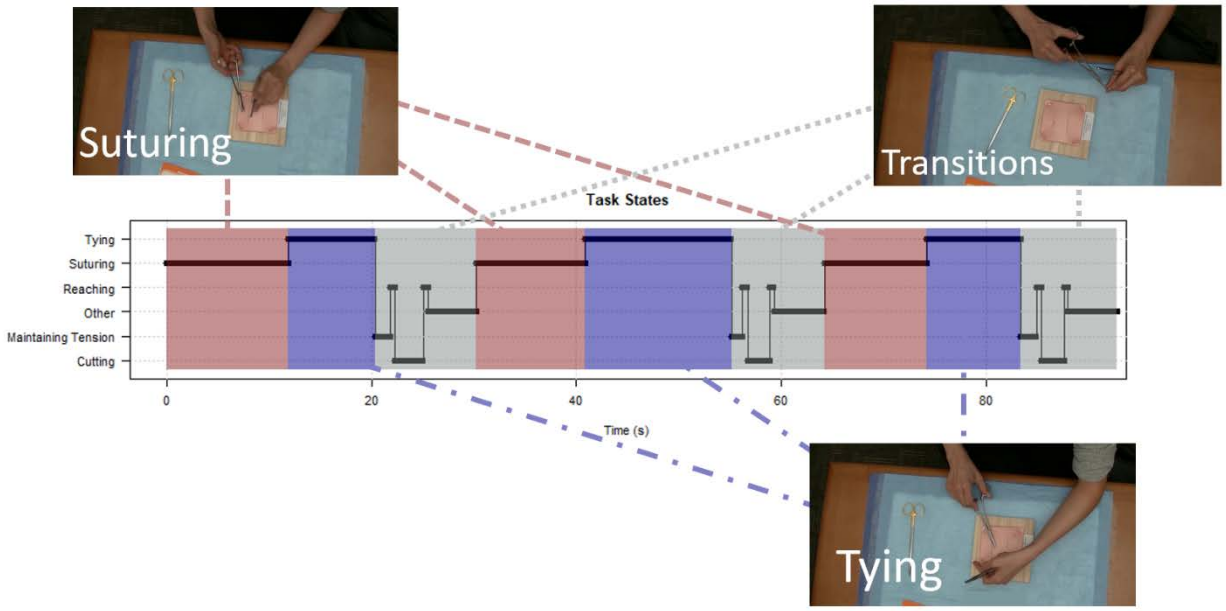


Figure 15: State representation of interrupted suturing task over time. The plot includes five state categories (e.g. tying, suturing, reaching, cutting, other) and the transitional "maintaining tension" state.

Each video was manually labeled using Multimedia Video Task Analysis (MVTA); a software platform developed at the University of Wisconsin-Madison (Yen & Radwin, 2007). Overall accuracy is determined by the percentage of periods (2s) classified correctly, while confusion matrices are presented to fully disclose the classification performance by task. Cycle estimations were calculated assuming five transitions within each transitional period, and completion of four knots in each tying period in order to complete the task.

3.3.4 Segmentation

To discretize the position record of the hands, successive spans of 60 frames (2 seconds) were chosen to encompass the lowest 5th percentile of tying tasks (mean = 5.1s, sd = 4.5s) and the minimum of suturing tasks (mean = 17s, sd =10s). These segments produced small periods to train the decision tree and decision tree ensemble. The classification was additionally improved through hidden markov modeling for simple interrupted suturing, in which the outputs from the decision tree ensemble provided emission symbols to the HMM. Running subcuticular suturing

state prediction was not tested in conjunction with HMMs, due to limited transitions between suturing and tying.

3.3.5 Machine Learning Approach

We tested three approaches to classify segments of surgical video into discrete states: decision trees, random forests, and HHMs. Decision trees are analogous to flow-charts, in which a series of if-statements are used to specify an outcome. Random forests, meanwhile, were used as an ensemble method (collection of decision trees) to decrease the variance in the prediction. HMMs were further employed to improve the random forest prediction by incorporating temporal transition information. HMMs were tested on a both a random subset of video segments, and a random set of participants. The random forest in each case provided the provided the observed emissions or symbolic input to the HMM, while the training data provided the transition probabilities. From these components, the HMM could predict a sequence of "likely states" for each of the testing cases.

Twenty percent of all video segments were randomly selected to serve as a test set for random forest prediction across all tasks, while thirty percent of participants from each experience group (medical students, junior residents, senior residents, attendings) comprised a testing set of twelve participants. This allowed for both random selection and within-user population accuracy prediction estimates.

3.4 Results

3.4.1 State Classification

Classification accuracy rates are presented in Table 5, with the best prediction models in bold. The cross-validation accuracy rate while training the random forest was 74%. The random selection of all participants (R) yielded similar accuracy (74%) on both the cross-validated training and testing data sets, while the participant-controlled approach (P) exhibited greater

difference between the training and testing sets, indicating some sensitivity to individual participant style or technique. Meaningful variables isolated every two seconds in the decision tree algorithm included: mean curvature values of the non-dominant hand, maximum distance between both hands, time the dominant hand spent within a radius of 22.5 centimeters of the simulation center, relative distance between the non-dominant hand and the simulation center, and the lateral path density of the non-dominant hand.

Table 5: Classification accuracy rates on testing set for each method. DT, decision tree; RF, random forest; CV, 10-fold cross validation accuracy; HMM, Hidden Markov Model; R, Random video segments across all participants; P, random selection of participants.

Segment Type	Decision Tree	Random Forest (CV)	RF + HMM*
Static (2s)	R: 0.64 P: 0.60	**R: 0.74 (0.74)** P: 0.68 (0.78)	R: 0.90 **P: 0.79**

*HMM are only applicable to simple interrupted suturing tasks.

The accuracy of RA+HMM for random selection (90%) is inflated because the HMM state prediction necessarily draws from time segments used both to train and test the random forest. The within-participant analysis (79% accuracy), on the other hand, predicts on a wholly reserved data partition. This is a better estimate of HMM improvement and thus shown in bold. The confusion matrix for the RF+HMM approach is presented in Table 6, while representative examples of good (> 90% accuracy) and poor (< 70% accuracy) state predictions are depicted as step plots in Figure 16.

Table 6: Confusion matrix for combined random forest and HMM classification (79% accuracy) approach on random subset of participants.

	Predicted		
Actual	Other	Suturing	Tying
Other	73.8% (330)	13.6% (64)	9.6% (43)
Suturing	5.8% (21)	75.1% (274)	19.2% (70)
Tying	8.2% (35)	6.3% (27)	85.5% (366)

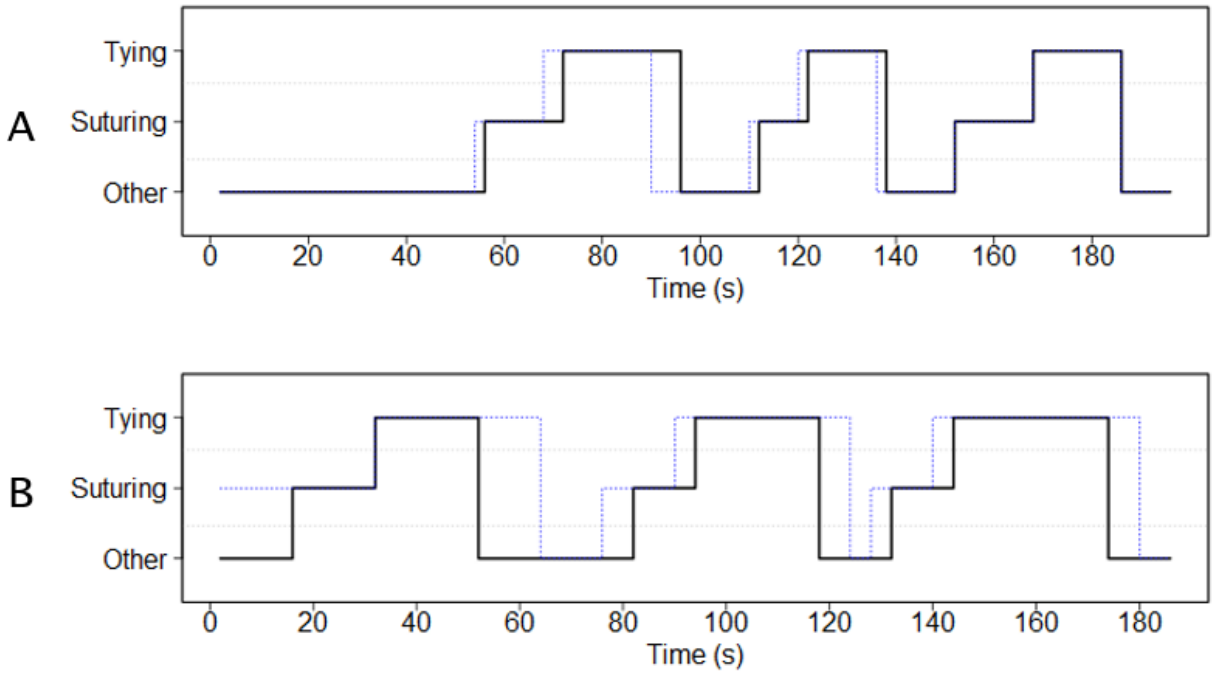


Figure 16: Representative examples of ground truth (solid) and predicted (dashed) good state prediction (A) with 91% classification accuracy, and poorer state prediction (B) with 70% classification accuracy.

3.4.2 Cycle Frequency

Inspection of Figure 16 indicates that even poorer models may still exhibit good sequence accuracy and be able to identify the number of cycles the rate of completion for repetitive tasks. Despite the small reserved sample size of participants, the HMM-derived cycle frequency

reasonably predicted the ground-truth labeled cycle rates on reserved test cases of at least two observed periods of suturing (Figure 17; slope = 0.88, intercept = 0.03, correlation = 0.83, R^2 = 0.72).

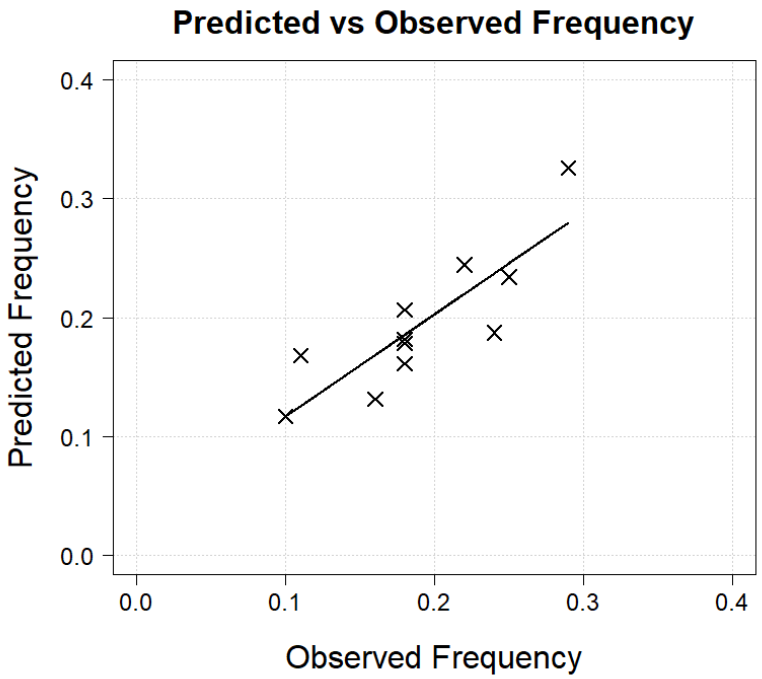


Figure 17: Predicted cycle frequency (hz) of combined tasks for reserved test cases.

3.5 Discussion

This study applied machine learning algorithms to classify surgical maneuvers and describe a participant's rate of progress through common benchtop suturing tasks. Three algorithms are implemented: decision trees, random forests, and Hidden Markov Models (HMMs). Random forest classification improved by HMMs yielded the best classification accuracy (79%) for a random subset of participants. The classification accuracy for the same approach on a random subset of time intervals across all participants approached 90%, however, this result is necessarily inflated by training data temporally interspersed with testing data. The classification schemes, like in other studies (Ahmidi et al., 2017), exhibited sensitivity to the

number of individual participants selected in the training data. Random sampling across all experience levels outperformed a subset of participants for each model. For the reserved cases, random forest prediction accuracy (68%) compared to the cross-validation accuracy (78%), is an indication of that the forest may be recognizing common, yet potentially inconsequential, attributes of behavior and over-fitting to the training set. Some participants, for instance, opted to cut with their non-dominant hand, rather than with their dominant hand, as most participants chose. Some participants also repositioned the tools mid-way through the experiment, introducing variability during periods of transition. Excluding these behaviors in the training set may drive mis-recognition.

Prasad and colleagues (2018), in a broad literature review on the implications of handedness among surgeons, identify concerns for non-dominant or hand-switching techniques, including the potential for needlesticks, and some evidence of increased complication rates. Still, the authors suggest that commitment to a selected technique, even if that technique deviates from the traditional approach, may outweigh only occasional “handed-appropriate action.” Our algorithms may be identifying similar patterns, as the locations and style of state transitions may be affected by these technique choices. Including such variation, to a reasonable degree, will be crucial for extrapolating state findings to other participants and settings.

The results of this study are consistent with a general 70-80% classification range for other maneuver-based classification studies (Reiley et al., 2008; Tao et al., 2013), but fall short of some surgeme level classifications such as “both hands pull” during knot tying, classified at close to 97% (Haro Bejar et al., 2012). Future work increasing the number of segments to 1 second or ½ second intervals and examining an increasing number of states may be able to incorporate these gesture-level movements and enable more direct comparison to more granular studies. However, part of the longer-term goal of this work is to facilitate retroactive digital

video review at a maneuver (or greater) level. Fast forwarding to portions of suspected suturing, for instance, would be of greater utility than identifying surgeme movements. Such a record would also better integrate with and support ongoing efforts to employ digital video as a coaching tool (Greenberg, Dombrowski, & Dimick, 2016; Y. Y. Hu et al., 2017; Soucisse et al., 2017). Still, surgeme or dexeme level quantification is contributing to building out the "ontological language of surgery," and continue to support quantitative novice-expert comparisons (French, Lendvay, Sweet, & Kowalewski, 2017). In parallel work (Azari, 2018), we have identified several kinematic features associated with changing status and tenure. Increasing the resolution of the state model in future work would allow for a more detailed comparison between maneuver based and surgeme based predictions for robotic and video-based surgical analysis.

This study predicts a state every two seconds, even though digital video is captured 30 times a second. Increasing the number of predictions per second or employing dynamic segmentations in future work may also yield greater flexibility and fine-tuning of start and end times of maneuvers. Previous explorations of hand motion signals captured with our tracking software suggest that the amplitude of the curvature signal may provide a reliable indicator of distinct motions (Akkas, Lee, Hu, Yen, & Radwin, 2016). In other words, local maxima in the curvature signal, accompanied by changes in speed, may indicate the beginning of a new movement. A movement-based segmentation function would depend on meeting three criteria: reaching some cumulative curvature angle value (A), over some distance (D), and spanning a minimal amount of time (T) (Figure 18), similar to that described by Beh and colleages (2011). The variation in arrival rates of dynamic segments, however, would render HMMs less appropriate for the prediction task, as a movement-determined transition rate from state to state would confound the probability of transition with the underlying hand movement metric.

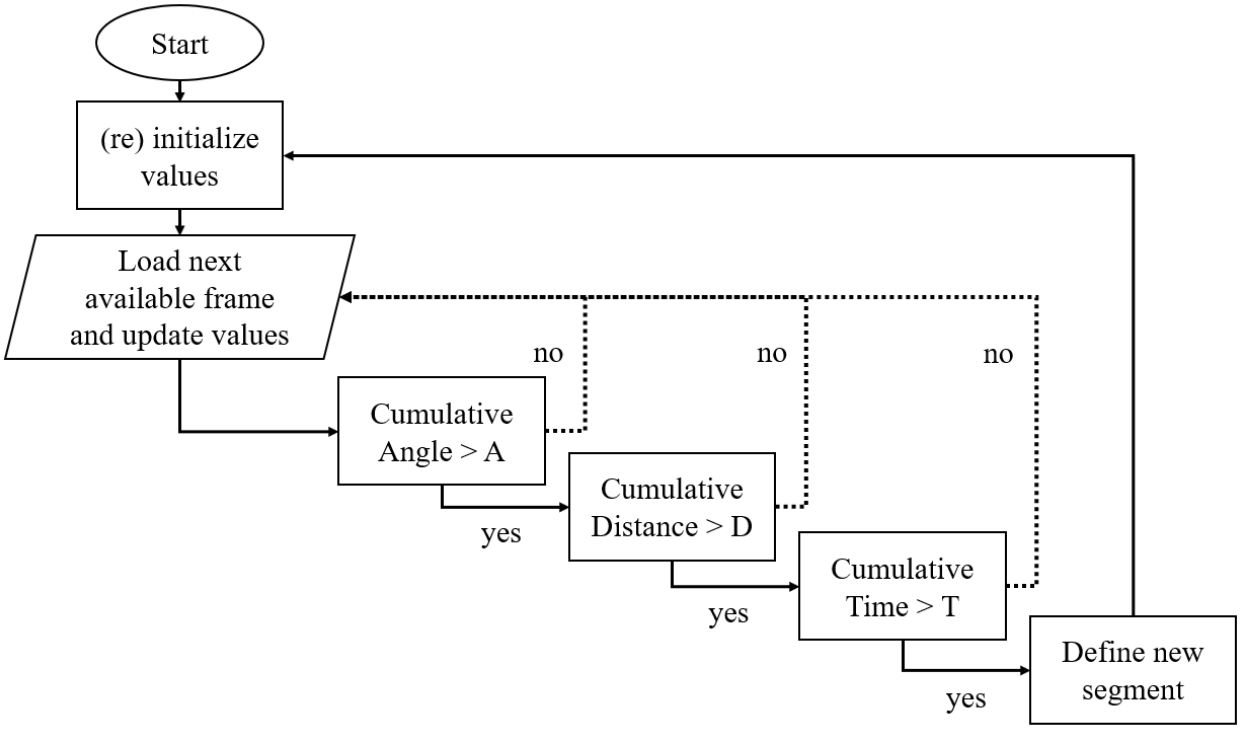


Figure 18: Flowchart of potential future, dynamic and trajectory-based segmentation function. New segments are defined if all three criteria are met: A, angle threshold; D, distance threshold; and T, time threshold.

Additional machine learning algorithms may address these limitations. Conditional Random Fields (CRF), in particular, are increasingly used to predict surgical gestures (Ahmidi et al., 2017; Dipietro et al., 2016; Elmezain, Al-Hamadi, & Michaelis, 2009; Sutton, 2012; Sutton & Mccallum, 2002; Sutton & McCallum, 2011), and may be more resilient to individual user style (Lea et al., 2015). CRF algorithms are built to optimize the conditional probability of states given all observations, rather than the joint probability of emissions while transitioning from state to state (i.e. markov process) optimized through HMM. There are a few existing software packages implemented across common programming platforms including Java, Python, R, Matlab, C# and C++ to assist in designing CRFs, but these sources lack the same level of maturing and customizability currently available to longer-studied models such as HMMs and random forests. Exploring the utility of custom CRF's in future work may improve classification

for marker less video tracking and dynamic segmentation functions, especially for longer and more variable task videos, such as running subcuticular suturing and porcine bowel anastomoses.

3.6 Conclusion

This study applied machine learning computer algorithms to automatically deconstruct surgical hand motion into discrete maneuvers. Random forest predictions improved through Hidden Markov Modeling achieved up to 90% accuracy on combined training and testing data, and 79% across experience levels on a reserved testing set. These results are similar to classification rates for robotic and laparoscopic studies for three-state models but fall short of current gesture-level classifications for distinct movements such as “both hands pulling”. Future directions for this work include increasing the number or flexibility of states employed to classify at the gesture and surgeme levels, implementing CRF prediction for digital video (thereby easing the reliance on transition probabilities), and extending random forest and HMM prediction to recorded video operating in different settings.

3.7 References

Aggarwal, R., Grantcharov, T., Moorthy, K., Milland, T., Papasavas, P., Dosis, A., … Darzi, A. (2007). An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Annals of Surgery, 245(6), 992–999. http://doi.org/10.1097/01.sla.0000262780.17950.e5

Ahmidi, N., Gao, Y., Béjar, B., Vedula, S. S., Khudanpur, S., Vidal, R., & Hager, G. D. (2013). String Motif-Based Description of Tool Motion for Detecting Skill and Gestures in Robotic Surgery. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 8149 LNCS, pp. 26–33). http://doi.org/10.1007/978-3-642-40811-3_4

Ahmidi, N., Poddar, P., Jones, J. D., Vedula, S. S., Ishii, L., Hager, G. D., & Ishii, M. (2015). Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. International Journal of Computer Assisted Radiology and Surgery, 10(6), 981–991. http://doi.org/10.1007/s11548-015-1194-1

Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B. B., … Hager, G. D. (2017). A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery. IEEE Transactions on Biomedical Engineering, 64(9), 2025–2041. http://doi.org/10.1109/TBME.2016.2647680

Akkas, O., Azari, D. P., Chen, C.-H. E. H. E., Hu, Y. H., Ulin, S. S., Armstrong, T. J., … Radwin, R. G. (2014). A hand speed – duty cycle equation for estimating the ACGIH hand activity level rating. Ergonomics, 58(2), 184–194. http://doi.org/10.1080/00140139.2014.966155

Akkas, O., Lee, C.-H., Hu, Y. H., Yen, T. Y., & Radwin, R. G. (2016). Measuring elemental time and duty cycle using automated video processing. Ergonomics, 59(11), 1514–1525. http://doi.org/10.1080/00140139.2016.1146347

Azari, D. P., Frasier, L. L., Quamme, S. R. P., Greenberg, C. C., Pugh, C. M., Greenberg, J. A., & Radwin, R. G. (2017). Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. Annals of Surgery, XX(Xx), 1. http://doi.org/10.1097/SLA.0000000000002478

Azari, D. P., Pugh, C. M., Laufer, S., Kwan, C., Chen, C. H., Yen, T. Y., … Radwin, R. G. (2015). Evaluation of Simulated Clinical Breast Exam Motion Patterns Using Marker-Less Video Tracking. Human Factors, 58(3), 427–440. http://doi.org/10.1177/0018720815613919

Barros, R. C., de Carvalho, A. C. P. L. ., & Freitas, A. A. (2015). Automatic Design of Decision-Tree Induction Algorithms. http://doi.org/10.1007/978-3-319-14231-9

Beh, J., Han, D., & Ko, H. (2011). Rule based trajectory segmentation applied to an HMM-based isolated hand gesture recognizer. Communications in Computer and Information Science, 174 CCIS(PART 2), 146–150. http://doi.org/10.1007/978-3-642-22095-1_30

Bhuyan, M. K., Ajay Kumar, D., MacDorman, K. F., & Iwahori, Y. (2014). A novel set of features for continuous hand gesture recognition. Journal on Multimodal User Interfaces, 8(4), 333–343. http://doi.org/10.1007/s12193-014-0165-0

Bouget, D., Allan, M., Stoyanov, D., & Jannin, P. (2017). Vision-based and marker-less surgical tool detection and tracking: a review of the literature. Medical Image Analysis, 35, 633–654. http://doi.org/10.1016/j.media.2016.09.003

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. http://doi.org/10.1023/A:1010933404324

Chen, C.-H., Azari, D. P., Hu, Y. H., Lindstrom, M. J., Thelen, D., Yen, T. Y., & Radwin, R. G. (2015). The accuracy of conventional 2D video for quantifying upper limb kinematics in repetitive motion occupational tasks. Ergonomics, 58(12), 2057–2066. http://doi.org/10.1080/00140139.2015.1051594

Chen, C.-H., Hu, Y. H., & Radwin, R. G. (2014). A motion tracking system for hand activity assessment. In 2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP) (pp. 320–324). IEEE. http://doi.org/10.1109/ChinaSIP.2014.6889256

Chen, C.-H., Hu, Y. H., Yen, T. Y., & Radwin, R. G. (2012). Automated Video Exposure Assessment of Repetitive Hand Activity Level for a Load Transfer Task. Human Factors:

The Journal of the Human Factors and Ergonomics Society, 55(2), 298–308. http://doi.org/10.1177/0018720812458121

Criminisi, A. (2011). Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. Foundations and Trends® in Computer Graphics and Vision, 7(2–3), 81–227. http://doi.org/10.1561/0600000035

Davids, K., Button, C., & Bennett, S. (2008). Dynamics of Skill Acquition - A Constraints-Led Approach. Champaign, IL, USA: Human Kinetics.

Despinoy, F., Bouget, D., Forestier, G., Penet, C., Zemiti, N., Poignet, P., & Jannin, P. (2016). Unsupervised Trajectory Segmentation for Surgical Gesture Recognition in Robotic Training. IEEE Transactions on Biomedical Engineering, 63(6), 1280–1291. http://doi.org/10.1109/TBME.2015.2493100

Dipietro, R., Lea, C., Malpani, A., Ahmidi, N., Vedula, S. S., Lee, G. I., … Hager, G. D. (2016). Recognizing surgical activities with recurrent neural networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9900 LNCS, 551–558. http://doi.org/10.1007/978-3-319-46720-7_64

Elmezain, M., Al-Hamadi, A., & Michaelis, B. (2009). Hand trajectory-based gesture spotting and recognition using HMM. Proceedings - International Conference on Image Processing, ICIP, 3577–3580. http://doi.org/10.1109/ICIP.2009.5414322

Fating, K., & Ghotkar, A. (2014). Performance Analysis of Chain Code Descriptor For Hand Shape Classification. International Journal of Computer Graphics & Animation (IJCGA), 4(2), 9–19.

Frasier, L. L., Azari, D. P., Ma, Y., Quamme, S. R. P., Radwin, R. G., Pugh, C. M., … Greenberg, C. C. (2016). A marker-less technique for measuring kinematics in the operating room. Surgery. http://doi.org/10.1016/j.surg.2016.05.004

French, A., Lendvay, T. S., Sweet, R. M., & Kowalewski, T. M. (2017). Predicting surgical skill from the first N seconds of a task: value over task time using the isogony principle. International Journal of Computer Assisted Radiology and Surgery, 12(7), 1161–1170. http://doi.org/10.1007/s11548-017-1606-5

Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2005). Mining data streams. ACM SIGMOD Record, 34(2), 18. http://doi.org/10.1145/1083784.1083789

Gao, Y., Vedula, S. S., Reiley, C. E., Ahmidi, N., Varadarajan, B., Lin, H. C., … Hager, G. D. (2014). JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling. Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop, 1–10.

Gavrila, D. . (1999). The Visual Analysis of Human Movement: A Survey. Computer Vision and Image Understanding, 73(1), 82–98. http://doi.org/10.1006/cviu.1998.0716

Glarner, C. E., Hu, Y. Y., Chen, C. H., Radwin, R. G., Zhao, Q., Craven, M. W., … Greenberg, C. C. (2014). Quantifying technical skills during open operations using video-based motion analysis. Surgery (United States), 156(3), 729–734. http://doi.org/10.1016/j.surg.2014.04.054

Greenberg, C. C., Dombrowski, J., & Dimick, J. B. (2016). Video-Based Surgical Coaching: An Emerging Approach to Performance Improvement. JAMA Surgery, 151(3), 282–3. http://doi.org/10.1001/jamasurg.2015.4442

Greene, R. L., Azari, D. P., Hu, Y. H., & Radwin, R. G. (2017). Visualizing stressful aspects of repetitive motion tasks and opportunities for ergonomic improvements using computer vision. Applied Ergonomics, 65, 461–472. http://doi.org/10.1016/j.apergo.2017.02.020

Haro Bejar, B., Zappella, L., & Vidal, R. (2012). Surgical gesture classification from video data. MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 7510(1), 34–41.

Hu, Y. Y., Mazer, L. M., Yule, S. J., Arriaga, A. F., Greenberg, C. C., Lipsitz, S. R., … Smink, D. S. (2017). Complementing operating room teaching with video-based coaching. JAMA Surgery, 152(4). http://doi.org/10.1001/jamasurg.2016.4619

Hundtofte, C. S., Hager, G. D., & Okamura, A. M. (2002). Building a task language for segmentation and recognition of user input to cooperative manipulation systems. Proceedings - 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, HAPTICS 2002, 225–230. http://doi.org/10.1109/HAPTIC.2002.998962

Iivarinen, J., & Visa, A. (1996). Shape Recognition of Irregular Objects. Intelligent Robots and Computer Vision XV: Algorithms, Techniques, Active Vision, and Materials Handling, 25–32.

Lea, C., Hager, G. D., & Vidal, R. (2015). An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks. Proceedings - 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015, 1123–1129. http://doi.org/10.1109/WACV.2015.154

Lea, C., Vidal, R., & Hager, G. D. (2016). Learning convolutional action primitives for fine-grained action recognition. In 2016 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1642–1649). IEEE. http://doi.org/10.1109/ICRA.2016.7487305

Lin, H. C. (2010). Structure in surgical motion. Johns Hopkins University.

Mackel, T., Rosen, J., & Pugh, C. (2007). Application of hidden markov modeling to objective medical skill evaluation. Studies in Health Technology and Informatics, 125, 316–318.

Maier-Hein, L., Vedula, S., Speidel, S., Navab, N., Kikinis, R., Park, A., … Jannin, P. (2017). Surgical Data Science: Enabling Next-Generation Surgery. Retrieved from https://arxiv.org/abs/1701.06482

Padoy, N., Blum, T., Ahmadi, S. A., Feussner, H., Berger, M. O., & Navab, N. (2012). Statistical modeling and recognition of surgical workflow. Medical Image Analysis, 16(3), 632–641. http://doi.org/10.1016/j.media.2010.10.001

Patel, V. L., Shortliffe, E. H., Stefanelli, M., Szolovits, P., Berthold, M. R., Bellazzi, R., & Abu-Hanna, A. (2009). The coming of age of artificial intelligence in medicine. Artificial Intelligence in Medicine, 46(1), 5–17. http://doi.org/10.1016/j.artmed.2008.07.017

Poppe, R. (2007). Vision-based human motion analysis: An overview. Computer Vision and Image Understanding, 108(1–2), 4–18. http://doi.org/10.1016/j.cviu.2006.10.016

Prasad, N. K., Kvasnovsky, C., Wise, E. S., & Kavic, S. M. (2018). The Right Way to Teach Left-Handed Residents: Strategies for Training by Right Handers. Journal of Surgical Education, 75(2), 271–277. http://doi.org/10.1016/j.jsurg.2017.07.004

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257–286. http://doi.org/10.1109/5.18626

Reiley, C. E., Lin, H. C., Varadarajan, B., Vagvolgyi, B., Khudanpur, S., Yuh, D. D., & Hager, G. D. (2008). Automatic recognition of surgical motions using statistical modeling for capturing variability. Studies in Health Technology and Informatics, 132(1), 396–401. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18391329

Rosen, J., Brown, J. D., Chang, L., Sinanan, M. N., & Hannaford, B. (2006). Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model. IEEE Transactions on Biomedical Engineering, 53(3), 399–413. http://doi.org/10.1109/TBME.2005.869771

Rosen, J., Hannaford, B., Richards, C. G., & Sinanan, M. N. (2001). Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE Transactions on Biomedical Engineering, 48(5), 579–591. http://doi.org/10.1109/10.918597

Soucisse, M. L., Boulva, K., Sideris, L., Drolet, P., Morin, M., & Dubé, P. (2017). Video Coaching as an Efficient Teaching Method for Surgical Residents—A Randomized Controlled Trial. Journal of Surgical Education, 74, 365–371. http://doi.org/10.1016/j.jsurg.2016.09.002

Starner, T., & Pentland, a. (1995). Real-time American Sign Language recognition from video using hidden Markov models. Computer Vision, 1995. Proceedings., International Symposium On, 265–270. http://doi.org/10.1109/ISCV.1995.477012

Sutton, C. (2012). An Introduction to Conditional Random Fields. Foundations and Trends® in Machine Learning, 4(4), 267–373. http://doi.org/10.1561/2200000013

Sutton, C., & Mccallum, A. (2002). An Introduction to Conditional Random Fields for Relational Learning. Graphical Models, 7, 93. http://doi.org/10.1677/JME-08-0087

Sutton, C., & McCallum, A. (2011). An Introduction to Conditional Random Fields. Machine Learning, 4(4), 267–373. http://doi.org/10.1561/2200000013

Tao, L., & Elhamifar, E. (2012). Sparse Hidden Markov Models for Surgical Gesture Classification and Skill Evaluation, 167–177.

Tao, L., Zappella, L., Hager, G. D., & Vidal, R. (2013). Surgical gesture segmentation and recognition. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8151 LNCS(PART 3), 339–346. http://doi.org/10.1007/978-3-642-40760-4_43

Uemura, M., Tomikawa, M., Kumashiro, R., Miao, T., Souzaki, R., Ieiri, S., … Hashizume, M. (2014). Analysis of hand motion differentiates expert and novice surgeons. Journal of Surgical Research, 188(1), 8–13. http://doi.org/10.1016/j.jss.2013.12.009

Vedula, S. S., Ishii, M., & Hager, G. D. (2017). Objective Assessment of Surgical Technical Skill and Competency in the Operating Room. Annual Review of Biomedical Engineering, 19(1), 301–325. http://doi.org/10.1146/annurev-bioeng-071516-044435

Vedula, S. S., Malpani, A., Ahmidi, N., Khudanpur, S., Hager, G., & Chen, C. C. G. (2016). Task-Level vs. Segment-Level Quantitative Metrics for Surgical Skill Assessment. Journal of Surgical Education, 73(3), 482–489. http://doi.org/10.1016/j.jsurg.2015.11.009

Wang, L., Hu, W., & Tan, T. (2003). Recent developments in human motion analysis. Pattern Recognition, 36(3), 585–601. http://doi.org/10.1016/S0031-3203(02)00100-0

Watson, R. A. (2014). Use of a Machine Learning Algorithm to Classify Expertise: Analysis of Hand Motion Patterns During a Simulated Surgical Task. Academic Medicine : Journal of the Association of American Medical Colleges, 89(8), 1–5. http://doi.org/10.1097/ACM.0000000000000316

Zappella, L., Béjar, B., Hager, G., & Vidal, R. (2013). Surgical gesture classification from video and kinematic data. Medical Image Analysis, 17(7), 732–745. http://doi.org/10.1016/j.media.2013.04.007

Zia, A., Zhang, C., Xiong, X., & Jarc, A. M. (2017). Temporal clustering of surgical activities in robot-assisted surgery. International Journal of Computer Assisted Radiology and Surgery, 12(7), 1171–1178. http://doi.org/10.1007/s11548-017-1600-y

Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). Hidden Markov Models for Time Series (2nd ed.). Boca Raton: CRC Press.

4. Automated video assessment for simulated surgical tasks of varying experience

4.0 Manuscript Information

This manuscript will be submitted to *Academic Medicine.*

4.1 Abstract

This study uses linear and generalized additive models of video-recorded hand motion to automatically predict expert-assessed surgical performance. Five experts rated anonymized video clips of benchtop suturing and tying tasks (n = 219) along four visual-analog (0-10) performance scales: motion economy, fluidity of motion, hand coordination, and tissue handling. Custom software enabled us to track the location of each of the recorded hand positions for all video frames and populate a robust feature set. A generalized additive model predicted fluidity of motion ratings with slope = 0.71, intercept = 1.98, and R^2 = 0.77. Fluidity of motion and motion economy models outperformed hand coordination and tissue handling. While hand motion tracking may not address all contextual features of surgical video, the kinematic features demonstrate that models of fluidity of motion and motion economy are generalizable across clinicians of different experience levels while suturing on foam. Future work will explore how well these simulation-based models extrapolate to the more dynamic settings of the operating room.

4.2 Background

There is increasing evidence that suggests improving skills promotes better patient outcomes (Birkmeyer et al., 2013). Surgical skill is traditionally comprised of technical and non-technical components (Yule et al., 2006; Steven Yule & Paterson-Brown, 2018). Madani and colleagues (2017) expanded these definitions through a broader conceptual framework of intraoperative performance. Using cognitive task analysis and literature review, the authors

identified five domains pertinent to surgical performance: 1) psychomotor skills, 2) declarative knowledge, 3) advanced cognitive skills, 4) interpersonal skills, and 5) personal resourcefulness. Surgeons must reach a competent level of performance in each of these domains lest they practice operating on live patients; an act for which, in the words of William Mayo is, “there’s no excuse” (Murphy, Torsher, & Dunn, 2007). Still, experience is gained while on the job: “you learn best from your mistakes, mistakes made on living people” (Groopman, 2008, p. 50).

Reflection of these difficult instances is considered a crucial component of expertise (Weinbergger et al., 2005). But during and after stressful situations, there is often little time to discuss; just when it is needed most (Pugh, 2014). Retrospective performance recall and feedback, a potential stop-gap for this issue, can “go stale” and lack reliability after just a few days (Bello et al., 2016, 2017). Judging competency and proficiency in these settings, as a result, still depends on direct, in-person, and real-time subjective observation. Without valid objective standards, Mieg (2009) found that socially determined ideas of professionalism like engagement and adherence to standards or educational involvement, may serve as stand-ins for more objective and quantified measures of skill and expertise, and what attributes comprise desirable performance.

The nearly “ubiquitous” (Newell & Rosenbloom, 1981) power law of learning, originally described by Snoddy (1926), and summarized by Stratton and colleagues (2007), is commonly applied to describe this performance in a series of skill acquisition stages (Anderson, 1982; Fitts & Posner, 1967). Movement through each of cognitive, associative and autonomous stages of skill acquisition are demonstrated by increased smoothness (Mohamadipanah et al., 2016), fluidity and automaticity (Crochet et al., 2011). Observable patterns like “slowing down” (Moulton et al., 2010), for instance, may indicate a transition from what Anderson, (1982) calls “autonomous” performance – invoking Kahneman's (2013) popularized “fast” or “automatic”

thinking as part of the dual processing theory (Evans & Stanovich, 2013) – to a more deliberate, cautious and state characterized by seeking for relevant cues, managing errors, and adjusting plans for several steps ahead. These kinds of cognitive transitions are a product of both surgeon expertise and case difficulty. A surgeon may employ great focused attention on the most complex aspects of a case, sometimes resulting in confusion during team changes and hand-offs (Wiegmann, Eggman, ElBardissi, Parker, & Sundt, 2010).

Davids and colleagues (2008) characterizes "skill acquisition as a learner (a dynamical movement system) searching for stable and functional states of coordination." While surgeons are required to demonstrate declarative knowledge, there is no comparable and widely accepted method to demonstrate operative dexterity. The best measure of psychomotor performance, the Objective Structured Assessment of Technical Skills (OSATS), is based on two criteria: (1) rating candidates along a series of Likert-based hand-motion scales and (2) tracking progress during a procedure on a specially-tailored checklist (Martin et al., 1997). Using the Spearman Brown prophecy formula (output = 2.15), the authors predicted that 8 stations would be needed to reach an 0.80 level of reliability across OSATS testing stations. Such expectations have largely borne out over time (Hatala, Cook, Brydges, & Hawkins, 2015). A meta-analysis of psychomotor skills assessments conducted by Jelovsek et al. (2013) found methodologically sound and well documented evidence for the OSATS Global Rating Scale (GRS) in accordance with the Accreditation Council for Graduate Medical Education (ACGME) guidelines. Additional studies examining OSATS have shown the prescribed assessment provides valid feedback as a formative assessment during coaching or training sessions (Hatala et al., 2015). Unfortunately, fully implementing OSATS has been described as requiring "effort and a budget outside that for daily medical practices" (Niitsu et al., 2013) and is generally considered too difficult to perform with regularity (Reznick, Regehr, MacRae, Martin, & McCulloch, 1997).

Objective computer-aided technical skill evaluation (OCASE-T) is of growing interest to expedite surgical skill assessment (Vedula et al., 2017). Sharma and colleages (2014), for example, have utilized frame kernel matrices and space time interest points (STIPS) to predict OSATS scores. Our group has previously collected open videos from the operating room for various tasks (Frasier et al., 2016) and predicted performance along a series of OSATS-derived scales from video-recorded hand motion (Azari et al., 2017).

There are also continued attempts to streamline assessment techniques, focusing on "shortcut assessments" (Hossein Mohamadipanah et al., 2018), "snap shot assessments" (Datta et al., 2006), text message rating schemes using the Zwisch performance scale (George et al., 2014), and 10-second classifications (French et al., 2017) based on the relationship between changing angle and speed of movement (Lacquaniti, Terzuolo, & Viviani, 1983). Crowd-sourced ratings, due to their expediency, are also being explored for dry lab suture tasks (White et al., 2014), laparoscopic procedures (Deal et al., 2016), and cricothyrotomy performed on a simulator (Aghdasi et al., 2015), among others. Vernez and colleagues (2017) found that applying crowd-sourced ratings of OSATS and Global Operative Assessment of Laparoscopic Skills (GOALS) (Gumbs, Hogle, & Fowler, 2007; Vassiliou et al., 2005) for laparoscopic skills "consistently identified top and bottom performers" in medical student populations seeking to enter residency. There is some evidence, however, that crowd-sourced ratings may be more lenient than expert ratings (Chen et al., 2014), and that individuals in the crowd, based on their reasoning, are not all equally accurate. The implications of using crowd-sourced assessment measures have yet to mature into actionable models of hands-on clinical performance. We seek to use expert generated performance ratings of short clips to develop a "gold standard" of hand-motion based performance measures across a spectrum of experience levels.

4.3 Hand Motion

Hand-motion patterns (i.e. kinematics) are of growing interest to objectively measure and predict a surgeon's skill before they are judged competent to operate. Studies have developed various metrics, ranging from motion density and movement rates (Azari et al., 2017), speed and acceleration (Frasier et al., 2016; Glarner et al., 2014), and signal entropy (Mackenzie, Watts, Patel, Yang, Hagegeorge, et al., 2016; Watson, 2014), to smoothness (Ghasemloonia et al., 2017), periods of idle time (D'Angelo, Rutherford, Ray, Laufer, et al., 2015), and total path length (Aggarwal et al., 2007), among others. Performance assessments in minimally invasive surgery (George, Skinner, Pugh, & Brand, 2018) and hands-on clinical palpation techniques (Laufer et al., 2016; Pugh, 2013) have been particularly amenable to instrumentation. Chmarra, Grimbergen, & Dankelman (2007) described 16 such tools and systems, the community of which has only grown since, to support what Maier-Hein and colleagues (2017) describe as the field of "surgical data science."

While many of these measures have intuitively and necessarily discriminated between experienced and novice performers, the binary "confirmation of such differences adds little" (Cook, 2015) to the overall validity argument (Kane, 2006, 2013). We hypothesize that in conjunction with expert rated performance along a continuous spectrum of experience, features of hand motion can form the basis of a performance model over the course of a career to predict and progression towards surgical proficiency and eventual decline. The goal of this study is to model performance ratings made by expert surgeons for a range of experience levels as participants complete two common benchtop suturing tasks.

This study builds on and extends preliminary work published by Azari, Frasier, et al., (2017). That study created regression models based on a series of kinematic features and used those models to predict subjective expert ratings of short segments of observed procedures. The current study extends this approach to test whether computer algorithms can discriminate

between subjective ratings for varying experiences. We hypothesize that the features of hand motion (synonymously referred to as hand movements or hand kinematics) can serve to model average expert ratings on observable performance for benchtop suturing tasks for varying experience levels.

4.4 Methods

4.4.1 Visual Analog Scales

We employed a series of subjective visual-analog rating scales (from 0 to 10) created and tested in a previous study (Azari et al., 2017). These scales include (A) fluidity of motion, (B) motion economy, (C) tissue handling and (D) hand coordination. These four scales were created to evaluate performance of short clips taken during live procedures of suturing and tying tasks. They are based upon existing OSATS global rating scales of instrument handling, time and motion, and respect for tissue. The goal of this study is to predict expert ratings along these scales of surgeons of various experience, as they complete common benchtop suturing tasks.

Fluidity of motion is a measure of hesitancy, pauses, or changes in direction and "resets," which may be a component of Moulton's "slowing down" (Moulton et al., 2010), contribute to time spent idle (D'Angelo, Rutherford, Ray, Laufer, et al., 2015), and represent the broader sensorimotor construct of "movement smoothness" (Balasubramanian, Melendez-Calderon, Roby-Brami, & Burdet, 2015). Motion economy is defined as efficiency of movement, or conservation of energy in any trajectory. Such behavior is consistently documented as a mark of expert psychomotor behavior (Davids et al., 2008) and has been suggested in creating surgical competency measures (Grober, Roberts, Shin, Mahdi, & Bacal, 2010). Tissue handling quantifies the appropriateness of the surgeon's force and tension when manipulating the tissue, and varies based on the tissue's friability (D'Angelo, Rutherford, Ray, Mason, et al., 2015; Laufer et al., 2016; Pugh, 2013). Coordination represents the simultaneous use of both hands – a potential

indicator of superior dexterity (Davids et al., 2008), and is reflected in the six domains in the Global Operative Assessment of Laparoscopic Skills (GOALS) (Gumbs et al., 2007; Vassiliou et al., 2005).

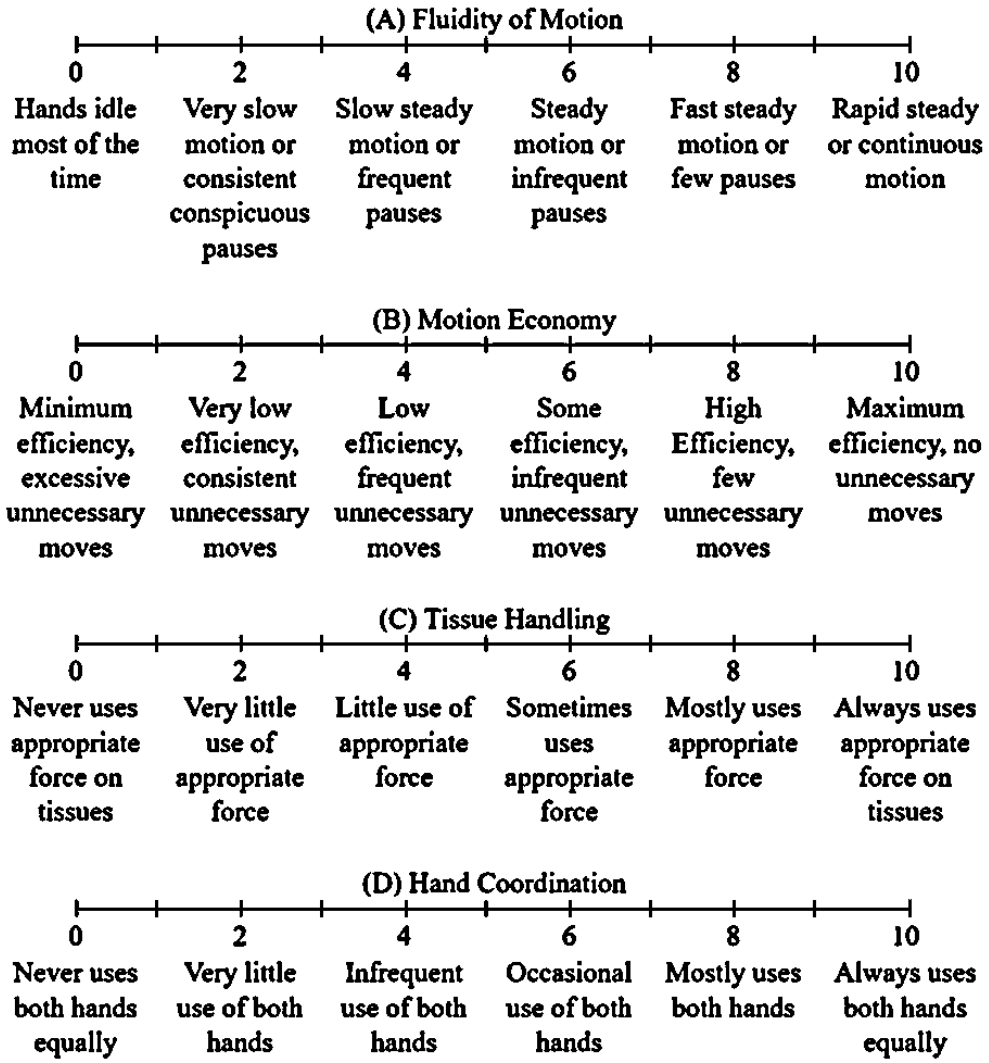


Figure 19: Visual analog performance rating scales (0-10) for expert review of tying and suturing video clips.

Two additional visual analog ratings scales: (E) independence, and (F) difficulty, adapted from the original GOALS (Vassiliou et al., 2005), were reserved for self-rating after the participants completed each task (Figure 20). Differences of self-ratings by task are described in parallel work by Azari et al., (2018, in press).

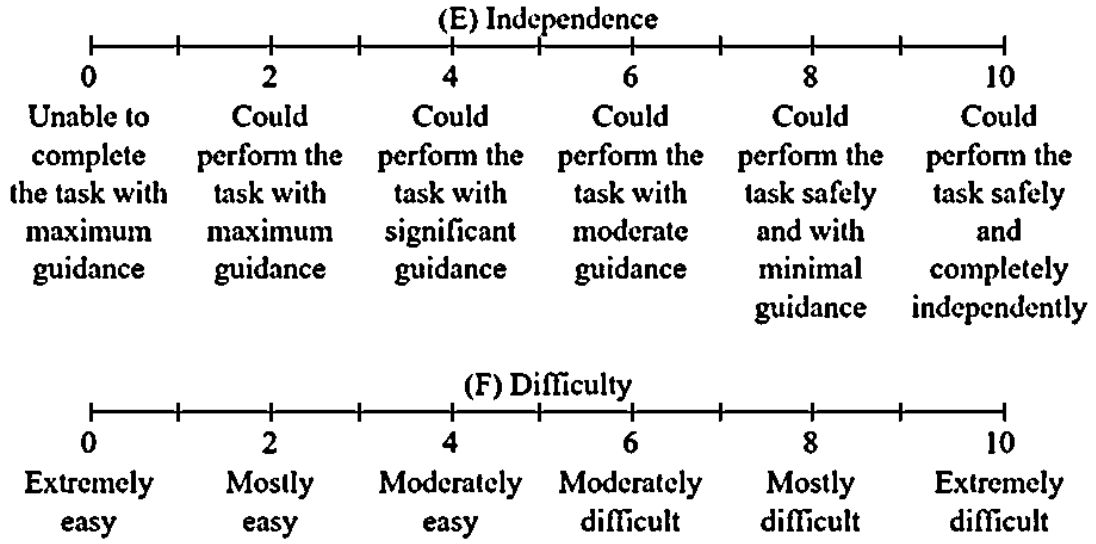


Figure 20: Visual analog scales for self-rating performance (0-10)

Procedure checklists were intentionally excluded from these rating schemes. The rating scales were designed to encompass the possible range of behavior (0-10). We have demonstrated the use of similar visual-analog rating scales to assess driver engagement and distraction (Radwin, Lee, & Akkas, 2017).

4.4.2 Participants and Setting

Thirty-seven participants enrolled in this study; each performing at least three simple interrupted stitches and a running subcuticular stitch of approximately 10 cm. Each participant's hand motions during the tasks were recorded and analyzed using digital video (Figure 21). Medical students (n=10) and surgical residents (n=15), were recruited to participate through announcements at the beginning of training and teaching sessions, while attending surgeons (n=10) and retired surgeons (n=2) were recruited through announcements at grand rounds, email list-serves and in-person discussion. Within the medical student population, six participants were in their fourth year, while with the resident population, ten "senior residents" had completed more than three post graduate years (PGY). Four of ten attending surgeons had six or more year of experience in their current position.

The University of Wisconsin Social and Behavioral Health and Science Institutional Review Board approved participant recruitment and participation. Prior to suturing, each participant reviewed and signed the consent agreement, and completed a short demographic survey to report their time in position, training status, and case load. Participation required approximately 10-15 minutes. While participant interaction was minimal, there were occasional periods where medical students would ask how to begin a task, or what kind of knot to use. These periods were manually identified and excluded from subsequent review.

Figure 21: Benchtop station (left) for anonymized video of suturing tasks (right).

Video cameras were positioned above each participant's working area to obtain a clear view of the working area while maintaining anonymity. Surgeon's faces were not visible. Cameras were activated only after signing the consent form, when the participant had the opportunity to ask questions. Notecards of known size allowed for video calibration.

The suturing tasks were simulated by two incisions – one for interrupted suturing and one for running subcuticular suturing. Incisions were cut in allevyn hydrocellular foam dressing (10.2 cm by 10.2 cm). The foam dressings were mounted to wood blocks (15 cm x 15 cm) so they would remain stable throughout the experiment. A small towel was placed under each dressing so that the foam would "pucker" and expose the interior of the incisions.

4.4.3 Motion Tracking

We used custom motion tracking software to record the position of the participants hands for each task. Written in C#, and using the OpenCVSharp vision libraries, this software can capable of recording the position of a region of interest (ROI) over the hand and track the position of the hand as it moves without sensors or markers. An analyst defines the initial ROI and monitors the software to ensure that any errors or jumps are manually corrected. In this study, the two-dimensional position of the hand was recorded every $1/30^{th}$ of a second, thereby enabling calculation of speed, acceleration, displacement, 2D-density, and path length measures, among others.

4.4.4 Video Review

Each video was examined for periods of between 20 and 80 seconds, in which several suturing cycles (i.e. stitches) were clearly visible. Initial cycles of medical student and resident suturing were treated as an adjustment period and omitted from expert review. Samples from simple interrupted suturing tasks (n=85) and running subcuticular tasks (n=134) comprised a dataset of 219 video clips, totaling 2 hours and 42 minutes of active suturing and tying.

Five expert surgeons independently rated each video clip from 0 to 10 along the four visual analog scales (Figure 19). Each panelist had at least three years of experience as an attending surgeon. The panelists viewed the clips and saved their scores via a software applet programmed in C# and distributed by USB (Figure 22). Raters completed a calibration activity to practice rating clips and compare their initial expectations to consensus scores from a previous study (Azari, Frasier, et al., 2017). Raters completed the activity at their convenience, and on different computers – saving progress over multiple sessions. Still, due to time constraints, not

all surgeons completed all ratings. Experts abstained from rating their own, performance, if applicable.

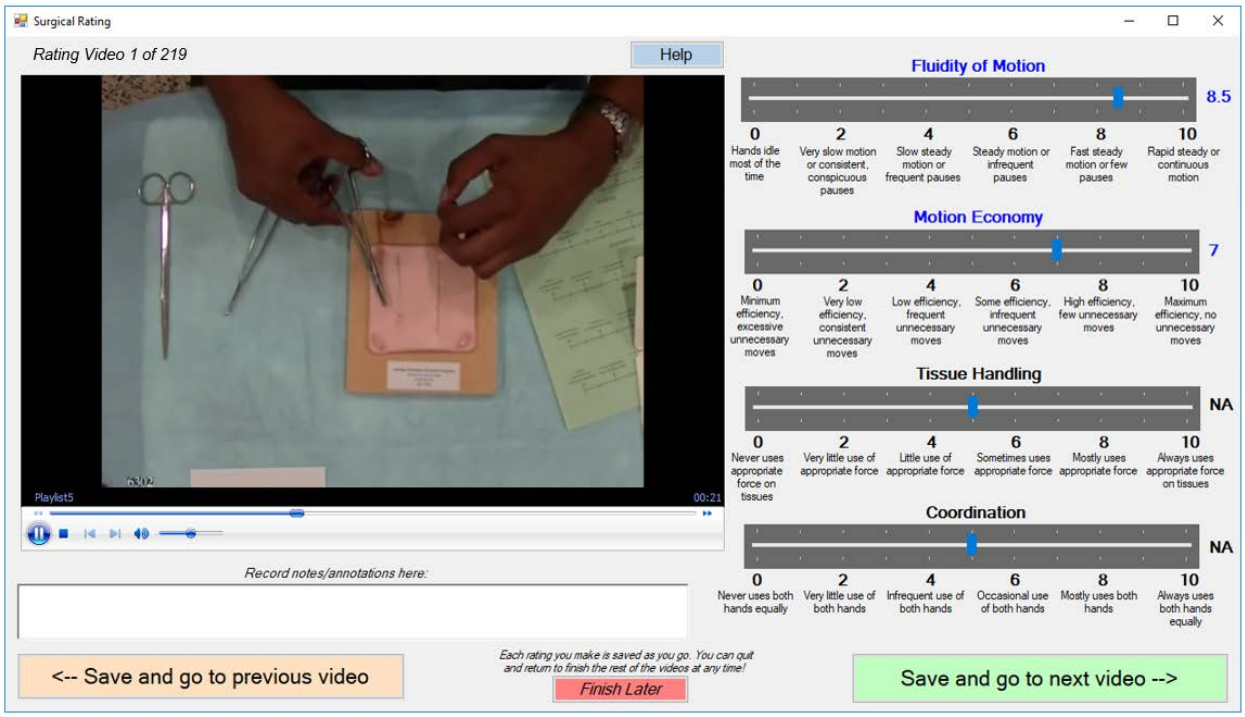


Figure 22: Rating applet showing auto-loaded video (upper-left) and visual analog scales (right). The first two visual analog scales have been manipulated to show that the interface provides active confirmation of all input ratings.

The average of judge ratings was considered appropriate for modeling if all ratings fell within a span of ±1.5 (range of 3). Clips where expert ratings spanned more than 3 were examined for outliers. If, after the removal of a suspected outlier, three of four total ratings fell within ±1 (range of 2), or four of five judges fell within ±1.25 (range of 2.5) the average of the remaining clip ratings were accepted for modeling. This approach is similar to existing methods, including randomly selecting of a subpanel of judges (Emerson & Arnold, 2011), or "mean-trimming," common in Olympic sport judgement, in which both high and low scores are removed (Emerson, Seltzer, & Lin, 2009; ISU, 2017). Our approach to outlier removal offers an additional advantage over many techniques, in that it avoids "outlier aversion" (J. Lee, 2004) by valuing both high and low ratings from each expert. This resembles the outlier detection

algorithm for academic judging proposed by Clemmensen et al. (2013), by examining the distribution of scores both before and after outlier removal. In this study, one outlier (at most) is removed if and only if the remaining panel ratings are sufficiently dense. The number of clips accepted for modeling along with the rates of clip acceptance by experience are shown in Table 7.

4.4.5 Rating Differences

Two non-parametric measures were used to examine interrater reliability: Krippendorff's alpha ($\propto$) and the average of pair-wise Spearman rank order correlation coefficients ($\bar{\rho}$). Krippendorff's alpha was selected for its ability to handle tied ranks and missing data efficiently (Krippendorff & Krippendorff, 2010), while the Spearman coefficient was selected due to its similarity to Kendall's W statistic as a non-parametric measure of concordance among several judges (Coleman & Coleman, 2015; Legendre, 2010; Olkin, Lou, Stokes, & Cao, 2015). The null hypothesis for Krippendorff's alpha is that rater agreement arises from chance. Tentative agreement measures are greater than 0.67, with better agreement at 0.80 or higher, and perfect agreement approaching 1.00. The Spearman rank correlation coefficient (ρ) also tests the null hypothesis that relationships between each pair of raters is random. Values of ρ greater than 0.50, accompanied by p values less than 0.05 indicate a low probability of randomness between the raters (Coleman & Coleman, 2015).

Interrater reliability (IRR) among panelists was also assessed using the intraclass correlation coefficient (ICC). ICC was applied as a two-way parametric comparison, assuming mixed-effects with multiple raters, adjusted to handle missing cases. Both consistency and absolute measures are reported. For more information on intraclass correlation coefficients and general assessment of interrater reliability, consult Koo and Li (2016) and Hallgren (2012), respectively.

Table 7: Number of ratings and percent (in parentheses) of clips meeting agreement criteria for each rating scale and participant experience level.

Experience	Years in Role	Fluidity of Motion	Motion Economy	Tissue Handling	Coordination
Junior Medical Student	≤ 3	16 (84%)	12 (63%)	12 (63%)	10 (53%)
Senior Medical Student	4	17 (65%)	18 (69%)	19 (73%)	19 (73%)
Junior Resident	< 3	18 (69%)	21 (81%)	20 (77%)	21 (81%)
Senior Resident	≥ 3	35 (70%)	40 (80%)	35 (70%)	44 (88%)
Junior Attending	< 6	29 (94%)	29 (94%)	28 (90%)	30 (97%)
Senior Attending	≥ 6	42 (78%)	44 (81%)	44 (81%)	44 (81%)
Retired	NA	6 (46%)	7 (54%)	8 (62%)	6 (46%)
Total		163 (74%)	171 (78%)	166 (76%)	174 (79%)

4.4.6 Modeling Process

A series of linear regression and grouped-lasso penalized generalized additive models (GAMS) were created, one for each rating scale, to predict the average ratings for each clip. GAMS represent each linear predictor as a sum of smoothed functions and enable less rigid relationship modeling (Chouldechova & Hastie, 2015). Features of tracked hand-motion served as independent variables and were examined for collinear relationships through standard Pearson correlations (Beysolow II, 2017; Kutner, Nachtsheim, Christopher, Neter, & Li, 2005). Parameters exhibiting variable inflation factors greater than 4 were also excluded. Skewed distributions of the response variables were transformed by subtracting each value from one greater than distribution maximum and applying a square root. Responses were transformed back to their original scale for plotting and comparison. Variables were selected utilizing penalized regression shrinkage methods including Ridge, LASSO, and Elastic Net (Hastie, Tibshirani, & Friedman, 2001). The number of parameters in each model was balanced against the mean-squared error (MSE) utilizing stepwise Akaike information criterion (AIC) selection (Akaike, 1974; Neter, Wasserman, & Kutner, 1990). Due to the complexity of generalized additive

models, the significance values associated with each coefficient in GAM models are associated with only linear components within the broader GAM. They are reported for comparison between variables but are generally considered therefore less reliable than those p-values provided for linear fits – a known phenomenon (J. D. Lee, Sun, Sun, & Taylor, 2016; Tibshirani, 2015), and may be improved in future techniques.

Judgement of model fit was contingent on plotting the predicted versus expert ratings, ideally appearing as a diagonal (0,0) to (10,10). The best models were arbitrarily defined to exhibit a slope between 0.5 and 1.5, with an intercept within ±2.5 of zero, and an $R^2 \geq 0.7$. Model validity was examined by comparing the leave-one-out predicted residual sum of squares (PRESS) measure, to the sum of squared errors (SSE), the same approach employed in a previous study examining surgical performance in the operating room (Azari et al., 2017). The PRESS measures were additionally converted to a predicted R^2 measure (R^2_{pred}) for ease of interpretation.

4.5 Results

4.5.1 Task Expert Rating Scales

In total, 219 video clips (mean time = 44 s, total time = 2.7 hrs) were individually rated by 5 attending surgeons along a series of visual analog scales. Raters observed 1476 active surgical cycles, including periods of suturing (n = 496), tying (n = 496), reaching (n = 181), cutting (n=139), and the transitional state between suturing and tying called maintaining tension (n = 177). Suturing comprised 60% of video clips, while tying comprised 24%. Less than 6% of observable time was classified as “other,” stemming for a small grace period at the beginning of each video clip for the rater to adjust to the new view. Video records of medical students yielded 45 clips from third year students (19 clips) and fourth year students (26 clips). Resident video yielded 76 clips, distributed between student in each of the post graduate years (PGY): PGY1 (11

clips), PGY2 (15 clips), PGY3 (15 clips), PGY4 (15 clips) and PGY5 (20 clips). Attending surgeon records yielded 85 samples, 54 of which were produced by attendings with more than 6 or more years tenure. Retired participants generated 13 clips.

Each scale was benchmarked from 0 to 10, intending to encompass the range of possible participant performance. Motion economy, fluidity of motion, tissue handling, and coordination were each rated for several cycles of simple interrupted and running subcuticular suturing. All expert ratings (n = 876) ranged from 0 to 10 for motion economy and coordination (mean = 6.1, 6.9, sd = 2.2, 2.0), and between 1 and 10 for fluidity and tissue handling (mean = 6.0, 7.0, sd = 2.3, 1.8). Ratings selected for modeling (n = 674) had similar means and standard deviations for each of motion economy (mean = 6.1, sd = 2.2), fluidity (mean = 6.1, sd = 2.3), tissue handling (mean = 7.2, sd = 1.7) and coordination (mean = 7.0, sd = 2.0), shown in Figure 23. Raters exhibited the greatest agreement for fluidity of motion both before and after outlier removal given Krippendorff's alpha ($\propto$ = 0.78, 0.81), and the average of the Spearman rank correlation coefficient ($\bar{\rho}$ = 0.75, 0.78). Motion economy exhibited good, but slightly reduced agreement ($\propto$ = 0.70, 0.75; $\bar{\rho}$ = 0.67, 0.73). There was less agreement for coordination ($\propto$ = 0.49, 0.56; $\bar{\rho}$ = 0.56, 0.57) and tissue handling ($\propto$ = 0.41, 0.52; $\bar{\rho}$ = 0.41, 0.51). The value of the Spearman rank correlation coefficient value $\bar{\rho}$ was significant ($p < 0.03$) for all scales but least meaningful for tissue handling ($p = 0.026$). Intraclass correlation coefficients (ICCs) between the means of all raters both before and after outlier removal are presented in Table 8.

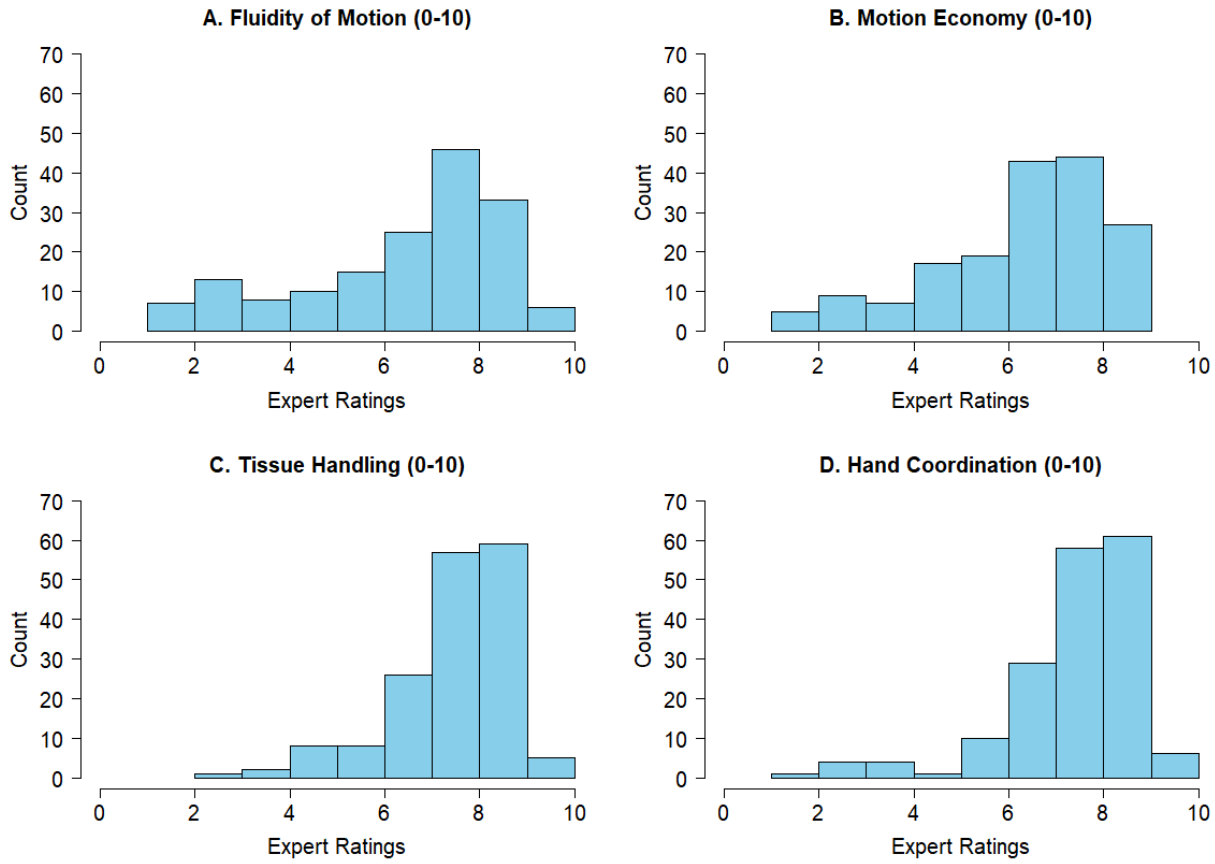


Figure 23: Distribution of expert ratings (0-10) selected for modeling for (A) fluidity of motion, (B) motion economy, (C) tissue handling and (D) hand coordination.

Table 8: Intraclass correlation coefficient (ICC) values (absolute / consistency) for each scale before (A) and after (B) outlier removal.

ICC Type		Fluidity of Motion	Motion Economy	Tissue Handling	Coordination
Single Raters	A	0.52 / 0.56	0.50 / 0.55	0.30 / 0.34	0.47 / 0.47
	B	0.65 / 0.68	0.63 / 0.66	0.40 / 0.45	0.49 / 0.55
Average Raters	A	0.84 / 0.86	0.84 / 0.86	0.69 / 0.72	0.78 / 0.82
	B	0.90 / 0.91	0.89 / 0.91	0.77 / 0.80	0.83 / 0.86

4.5.2 Prediction Models of Expert Ratings

Variables and significance values for predictors in each model are included in Table 9.

Fluidity of Motion

A GAM to predict fluidity of motion ratings (Figure 24A; slope = 0.71, intercept = 1.98, R^2 = 0.77, correlation = 0.88, R^2_{pred} = 0.70) provided a generally better fit than the linear model (slope = 0.67, intercept = 2.21, R^2 = 0.69, correlation = 0.83, R^2_{pred} = 0.68), with similar R^2_{pred} values.

Motion Economy

The linear model for motion economy (Figure 24B; slope = 0.65, intercept = 2.36, R^2 = 0.66, correlation = 0.81) exhibited small sensitivity to individual cases (R^2_{pred} = 0.61). A GAM for motion economy improved the R^2 value (slope = 0.76, intercept = 1.68, R^2 = 0.76, correlation = 0.87) but sacrificed generalizability by increasing reliance on individual case performance (R^2_{pred} = 0.59).

Tissue Handling

A linear model for tissue handling (Figure 24C; slope = 0.52, intercept = 3.65, R^2 = 0.57, correlation = 0.75, R^2_{pred} = 0.50) performed slightly better than its GAM counterpart (slope = 0.45, intercept = 4.19, R^2 = 0.54, correlation = 0.74, R^2_{pred} = 0.49). The slopes and intercepts for these predictions were not substantially different for those clips rated greater than five, despite the lower density of ratings.

Hand Coordination

The GAM to predict hand coordination rating (Figure 24D; slope = 0.55, intercept = 3.40, R^2 = 0.63, correlation = 0.79, R^2_{pred} = 0.44) provided a slightly better fit than the linear model (slope = 0.43, intercept = 4.33, R^2 = 0.46, correlation = 0.68, R^2_{pred} = 0.42), but both versions exhibited sensitivity to individual records as evident in their low R^2 and R^2_{pred} values.

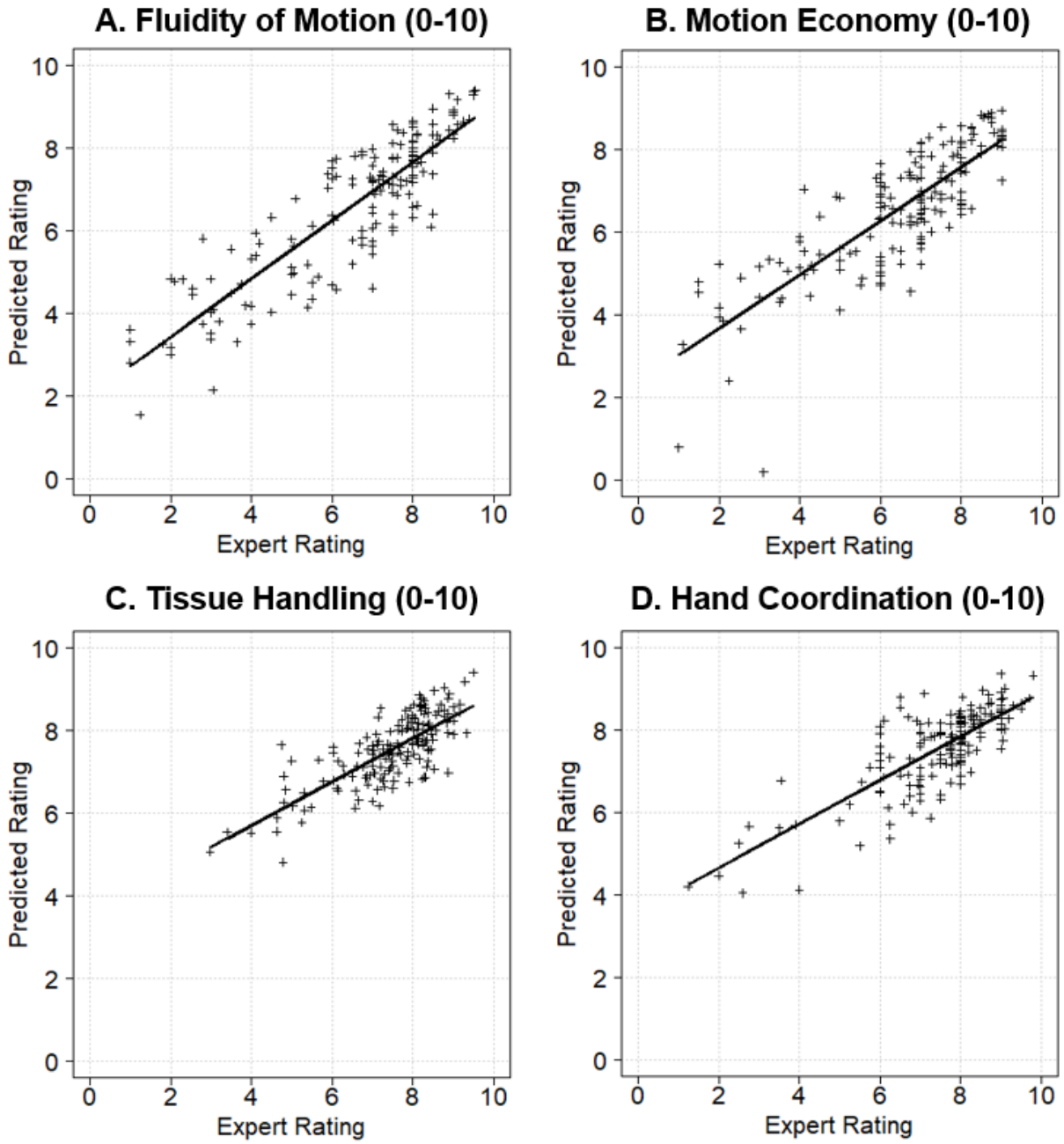


Figure 24: Predicted ratings vs the observed expert ratings for rating scales (A) fluidity of motion, (B) motion economy, (C) tissue handling, and (D) hand coordination.

Table 9: Regression model summary statistics and variables. GAM = generalized additive model; LM = linear model; Pred = predicted, Obs = observed, m = slope, b = intercept, with predicted = m * (observed) + b; R^2 = coefficient of determination; R^2_{pred} = PRESS statistic derived coefficient of determination; FFT, Fast Fourier Transform, D indicates dominant hand; N, non-dominant hand; B, combined hands; F, frames; T, threshold in in mm for distance, mm/s for speed, mm/s^2 for acceleration, counts for path densities.

Fluidity of Motion (GAM)

Pred vs Obs	Variables	p-value
m = 0.71	Cycle rate (hz)	0.000
b = 2.01	Curvature signal peak variance (N, T = 0.5)	0.005
c = 0.88	Minimum forward distance from center (D)	0.012
R^2 = 0.76	Minimum (5%) distance between hands	0.000
R^2_{pred} = 0.69	Peak arrival rate in speed signal (N, T = 250)	0.189
	Time (F) spent within center radius (N, T = 175)	0.000
	FFT frequency component of acceleration (D)	0.000
	Time (%) of both hands in motion above speed 75 mm/s	0.053
	Peaks in curvature signal (D, T = 0.5)	0.001
	Lateral path density (N, 40 mm sections, F = 180)	0.000

Motion Economy (LM)

Pred vs Obs	Variables	p-value
m = 0.65	Cycle rate (hz)	0.000
b = 2.36	Curvature signal peaks (D, T = 0.5)	0.005
c = 0.81	Curvature signal peaks (N, T = 0.0)	0.003
R^2 = 0.66	Smoothed speed signal peaks (N, T = 100)	0.024
R^2_{pred} = 0.61	Minimum (5%) distance between hands	0.109
	Standard deviation of distance between hands	0.091
	Peak arrival rate of speed (N, T = 200)	0.005
	Minimum forward distance from center (D)	0.158
	Time (F) spent within center radius (N, T = 150)	0.000
	Time (%) of both hands moving above speed 75 mm/s	0.018
	Lateral path density (N, 40mm range, F = 90)	0.001
	Path length per cycle (B)	0.029

Tissue Handling (LM)

Pred vs Obs	Variables	p-value
m = 0.52	Cycle rate (hz)	0.000
b = 3.65	Curvature signal peaks (D, T = 0.5)	0.000
c = 0.75	Curvature signal peak variance (N)	0.031
R^2 = 0.57	Standard deviation of distance center (D)	0.000
R^2_{pred} = 0.50	Time (%) within radius of center (D, T = 125)	0.005
	FFT frequency component of acceleration (D)	0.002
	Path density (N, 40 mm range, T = 30)	0.000

Hand Coordination (GAM)

Pred vs Obs	Variables	p-value
m = 0.53	Cycle rate (hz)	0.000
b = 3.60	Curvature signal peak variance (N)	0.093
c = 0.77	Smoothed acceleration peaks (N)	0.000
R^2 = 0.60	Minimum forward distance from center (D)	0.205
R^2_{pred} = 0.44	FFT frequency component of acceleration (D)	0.125

4.6 Discussion

This study uses digital video and motion tracking of clinician hand movements to predict expert ratings along four subjective scales: fluidity of motion, motion economy, tissue handling, and hand coordination. The scales are adapted for short clips from the existing global motion scales of OSATS and GOALS and designed to encompass the full range of participant behavior. While a broad selection of scores for each scale were generated, the scores were not uniformly distributed. All scales showed average ratings between 6 and 8, with tissue handling, and hand coordination ratings appearing most skewed (Figure 23). Expert ratings for fluidity of motion and motion economy were the most concordant, as judged by ICC and the non-parametric Kendall's W.

Two modeling approaches were employed: ordinary least squares linear regression with AIC stepwise reduction and generalized additive modeling (GAM). Multicollinear variables were removed utilizing pairwise elimination, variable inflation factor analysis, and a series of penalized regression shrinkage methods. The generalized additive model performed best in predicting fluidity of motion ratings, while a linear model best predicted expert ratings best for motion economy, despite some underprediction for expert ratings of 4 or less (see Figure 24B). Tissue handling predictions and hand coordination predictions underperformed relative to motion economy and fluidity of motion, but consistently predicted performance ratings of 7 and higher.

The GAM for fluidity of motion offered the best prediction results, with over 95% of residuals falling within ± 1.5. The fluidity ratings differed significantly between medical students (mean = 4.1, sd = 1.9) and all other groups ($p < 0.04$), as well as between senior (mean = 6.8, sd = 1.5) and junior (mean = 5.7, sd = 1.6) residents ($p < 0.03$). These differences support generalization of this model within a testing environment. It is important to note that the

significance of each of the predictors in the GAM were reported using only their linear components and is a general limitation of this technique.

There were several variables common to each of the prediction functions. Cycle frequency (Hz), peak arrival in the curvature signal, as well as the FFT frequency transform of acceleration of the dominant hand were common predictors across each of the sets. Similarly, the peak arrival rates in both the filtered and unfiltered speed and acceleration signals repeatedly appeared to be significant factors in predicting fluidity and motion economy. Positional relationships between both hands also held an important role in fluidity of motion, motion economy, and tissue handling. Consistent significance of both dominant and non-dominant hand movements reflect findings in previous studies: that more experienced surgeons leverage activity of the non-dominant hand to improve overall efficiency (Glarner et al., 2014).

Given the central role of cycle rates in predicting these performance ratings, it is crucial to note that the cycle rates were determined manually for these tasks – a significant caveat. However, accompanying work by (Azari, 2018) has developed automatic recognition and prediction of observed cycle rates, enabling video-based calculation of cycle frequency for simple benchtop tasks. Future work will need to integrate the cycle rates into the prediction models.

Despite the progress made predicting fluidity of motion and motion economy, this study is limited to performance predictions that are only proxy measures of skill, and do not account for contextual variations in the surgical setting, team, and overall case complexity – necessary considerations to reduce surgical errors more broadly as part of a work-systems based approach (Wiegmann et al., 2010). Similarly, this approach to motion tracking, despite good predictions for fluidity of motion and motion economy, cannot assess the quality of the completed maneuver, or ensure that the maneuver is complete. In comparing two trajectories, for instance – one in

which a clinician is holding a ligature and completing a suturing task, and one where the clinician is demonstrating the hand motion of suturing while empty-handed – may look identical to the computer. A rater, however, may recognize the context of these two cases and discount early empty-handed movements, and adjust their rating. Participants may also reposition themselves to gain better leverage during a task, changing the location density of the hands, but not necessarily impact the rating score. We suspect that these kinds of contextual inferences are driving the reduced R^2 value and outliers in the motion economy prediction scheme, in which one of the greater outlier cases (expert rating of 3) consists of a participant repeatedly changing positions to gain a stable posture before driving a needle through tissue. These examples are important to include to paint a complete picture of how the interface handles less-prevalent behaviors, or behaviors over significantly different time periods than provided in the training data.

Tissue handling ratings may depend on similar contextual inferences. When describing tissue handling, surgeons may address how tissue responds to changing force (e.g. “the skin puckers”, or “raises up”), and an improved sensitivity along a finger pad to help to fine tune the amount of force applied on sensitive tissue, rather than larger muscle movements initiated from the shoulder or forearm applying additional leverage during instrument tying. These factors are not readily observed by our tracking routine, but their impact in real-world situations represents a substantial risk. Knowing that the tissue in these videos was foam, may have rendered this study less applicable to a stable prediction of tissue handling ratings. Tissue handling predictions in previous studies have also underperformed relative to fluidity of motion and motion economy rating predictions (Azari et al., 2017). In both studies, the range of observed tissue handling scores was also reduced compared to fluidity of motion and motion economy. Tissue handling may be more difficult to predict from two-dimensional hand position records but may also be

more difficult to assess from limited cues available during short clips of video. Discerning forces from video is an ongoing effort in a variety of industrial engineering applications. Future study is needed to solidify the relationship between video-derived hand motion records and contextual adaptations and forces central to respecting tissue. Additional studies may benefit by comparing performance recorded on both video-based and senor-integrated platforms.

Hand coordination ratings, meanwhile, may be affected by a clinician preference of how best to alternate use of dominant and non-dominant hands. Previous work by our group has identified significant differences in non-dominant hand movement and displacement commensurate with clinician experience (Glarner et al., 2014). There is continuing discussion on how to balance workload between dominant and non-dominant hands (Burdett, Dunning, Goodwin, Theakston, & Kendall, 2016; Prasad et al., 2018), and on training implications of left-handed surgeons in general (M. Anderson, Carballo, Hughes, Behrer, & Reddy, 2017; Dobson, 2005).

The underlying premise of hand coordination is that sharing workload between the hands can maximize efficiency. This may manifest differently by chosen technique, or by experience. Attendings, for instance, accelerated their non-dominant hand less frequently than senior residents, despite completing the task faster utilizing both hands (Azari, 2018). One-handed or two-handed tying in the operating room could also prompt different hand coordination scores, despite a similar task outcome. Although unavailable for this study, future work may express hand coordination as active engagement by both hands or passing materials between hands, as it has been in laparoscopic studies (Law et al., 2016), and inferred during periods of reduced speed during simulated clinical breast exams (Azari et al., 2015). While it is possible that the motion economy ratings may already encompass hand coordination as a constituent, future research is

needed to determine the extent to which effective hand coordination depends on unevenly balancing work between hands.

There were many challenges in completing this study. Rating collection, due to the valuable time experts devoted, was the most laborious. A custom rating applet was programmed and distributed on USB to ease the burden on experienced clinicians. Raters could complete the experiment at their convenience over multiple periods. Over the course of the study, the program was improved to allow rating on different computers, and better control over multiple sessions. Rating each clip, however, still required several hours, and comprised the most substantial portion of this work. This challenge clarifies the underlying difficulty in implementing OSATS more broadly and highlights the future utility automated feedback routines.

Recent excitement surrounding crowdsourcing to predict performance of laparoscopic video (Vernez et al., 2017) suggests that large numbers of inexperienced viewers may be a reliable alternative to collecting expert rating performance. Such an approach, however, is not without concerns. Reliance on expert opinion – especially individuals in positions to grade and assess potential students, grounds expectation of surgical performance within existing uses, and casts less doubt on the source of scores. Despite the additional time and effort, the expert ratings in this study provide an excellent backdrop from which to quantify surgical motion without relying on large numbers of untrained eyes.

Supervision of the software platform while motion tracking also posed a substantial burden – primarily to identify and manually control for periods of out-of-frame motion or unexpected occlusions. A five-minute video clip required between 20 and 60 minutes to track both participant hands effectively. Continuing improvements in motion tracking algorithms, in conjunction with greater interface control over tracking parameters (currently under development) will ease this burden in the future.

4.7 Conclusion

This study created prediction models of expert-rated performance assessments (0-10) for clinicians of varying experience completing common benchtop suturing tasks. The best prediction model was achieved for fluidity of motion (slope = 0.71, intercept = 2.01, R^2_{pred} = 0.69). Several variables were significant predictors across each scale, including the cycle frequency, the peak arrival rate in the speed and acceleration signals, and the main frequency component of the FFT for acceleration of the dominant hand. While cycle frequency is manually calculated for the current study, the subsequent chapter describes increasing success in automatically predicting cycle rates. The prediction functions created in this study, if packaged in a stand-alone application, could provide active feedback scores to medical students and residents hoping to improve their performance along each scale, and gain a general understanding of their current surgical dexterity. Considering Kane and Messick's modern validity framework, the intended use of these prediction functions would be to offer a general suite of scores to augment formative feedback. They should not usurp or take the place of an experienced coach.

4.8 References

Aggarwal, R., Grantcharov, T., Moorthy, K., Milland, T., Papasavas, P., Dosis, A., … Darzi, A. (2007). An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Annals of Surgery, 245(6), 992–999. http://doi.org/10.1097/01.sla.0000262780.17950.e5

Aghdasi, N., Bly, R., White, L. W., Hannaford, B., Moe, K., & Lendvay, T. S. (2015). Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. The Journal of Surgical Research, 196(2), 302–6. http://doi.org/10.1016/j.jss.2015.03.018

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723. http://doi.org/10.1109/TAC.1974.1100705

Anderson, J. R. (1982). Acquisition of Cognitive Skill. Psychological Review, 89(4), 369–406. http://doi.org/10.1037/0033-295X.89.4.369

Anderson, M., Carballo, E., Hughes, D., Behrer, C., & Reddy, R. M. (2017). Challenges training left-handed surgeons. American Journal of Surgery, 214(3), 554–557. http://doi.org/10.1016/j.amjsurg.2016.12.011

Azari, D. P. (2018). Quantifying Surgical Skill. University of Wisconsin-Madison.

Azari, D. P., Frasier, L. L., Quamme, S. R. P., Greenberg, C. C., Pugh, C. M., Greenberg, J. A., & Radwin, R. G. (2017). Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. Annals of Surgery, XX(Xx), 1. http://doi.org/10.1097/SLA.0000000000002478

Azari, D. P., Greenberg, J. A., Miller, B. L., Le, B. V., Greenberg, C. C., Pugh, C. M., … Radwin, R. G. (2018). Can Surgical Performance for Varying Experience be Measured from Hand Motion? [in press]. In 2018 Annual Meeting of the Human Factors and Ergonomics Society Conference Proceedings (p. 5).

Azari, D. P., Pugh, C. M., Laufer, S., Kwan, C., Chen, C. H., Yen, T. Y., … Radwin, R. G. (2015). Evaluation of Simulated Clinical Breast Exam Motion Patterns Using Marker-Less Video Tracking. Human Factors, 58(3), 427–440. http://doi.org/10.1177/0018720815613919

Balasubramanian, S., Melendez-Calderon, A., Roby-Brami, A., & Burdet, E. (2015). On the analysis of movement smoothness. Journal of Neuroengineering and Rehabilitation, 12(1), 112. http://doi.org/10.1186/s12984-015-0090-9

Bello, R. J., Meyer, M. L., Cooney, D. S., Rosson, G. D., Lifchez, S. D., & Cooney, C. M. (2017). Reliability of Operative Skill Evaluations: How Late Is Too Late to Give Feedback? Plastic and Reconstructive Surgery - Global Open, 5(9), e1465. http://doi.org/10.1097/GOX.0000000000001465

Bello, R. J., Sarmiento, S., Rosson, G. D., Cooney, D. S., Lifchez, S. D., & Cooney, C. M. (2016). Understanding the Role for Operative Performance Rating Tools in Meeting Surgical Trainee Feedback Needs. Plastic and Reconstructive Surgery - Global Open, 4(6), e780. http://doi.org/10.1097/GOX.0000000000000777

Beysolow II, T. (2017). Introduction to Deep Learning Using R. Berkeley, CA: Apress. http://doi.org/10.1007/978-1-4842-2734-3

Birkmeyer, J. D., Finks, J. F., O'Reilly, A., Oerline, M., Carlin, A. M., Nunn, A. R., … Birkmeyer, N. J. O. (2013). Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine, 369(15), 1434–42. http://doi.org/10.1056/NEJMsa1300625

Burdett, C., Dunning, J., Goodwin, A., Theakston, M., & Kendall, S. (2016). Left-handed cardiac surgery: Tips from set up to closure for trainees and their trainers. Journal of Cardiothoracic Surgery, 11(1), 1–4. http://doi.org/10.1186/s13019-016-0523-y

Chen, C., White, L., Kowalewski, T., Aggarwal, R., Lintott, C., Comstock, B., … Lendvay, T. (2014). Crowd-Sourced Assessment of Technical Skills: a novel method to evaluate

surgical performance. Journal of Surgical Research, 187(1), 65–71. http://doi.org/10.1016/j.jss.2013.09.024

Chmarra, M. K., Grimbergen, C. a, & Dankelman, J. (2007). Systems for tracking minimally invasive surgical instruments. Minimally Invasive Therapy & Allied Technologies : MITAT : Official Journal of the Society for Minimally Invasive Therapy, 16(6), 328–40. http://doi.org/10.1080/13645700701702135

Chouldechova, A., & Hastie, T. (2015). Generalized Additive Model Selection. Annals of Applied Statistics, 1–24. Retrieved from http://arxiv.org/abs/1506.03850

Clemmensen, Line Harder; Rosas, Harvey; Thompson, M. K. (2013). Statistical Outlier Detection for Jury Based Grading Systems. 120th Asee Annual Conference and Exhibition, 1–14. Retrieved from http://orbit.dtu.dk/files/54060539/Statistical_Outlier_Detection.pdf

Coleman, J. S. M., & Coleman. (2015). Spearman Rank Order Correlation. In N. J. Salkind (Ed.), Encyclopedia of Research Design (pp. 1405–1408). Thousand Oaks, CA: SAGE Publications, Inc.

Cook, D. A. (2015). Much ado about differences: why expert-novice comparisons add little to the validity argument. Advances in Health Sciences Education : Theory and Practice, 20(3), 829–34. http://doi.org/10.1007/s10459-014-9551-3

Crochet, P., Aggarwal, R., Dubb, S. S., Ziprin, P., Rajaretnam, N., Grantcharov, T., … Darzi, A. (2011). Deliberate practice on a virtual reality laparoscopic simulator enhances the quality of surgical technical skills. Annals of Surgery, 253(6), 1216–1222. http://doi.org/10.1097/SLA.0b013e3182197016

D'Angelo, A. L. D., Rutherford, D. N., Ray, R. D., Laufer, S., Kwan, C., Cohen, E. R., … Pugh, C. M. (2015). Idle time: An underdeveloped performance metric for assessing surgical skill. American Journal of Surgery, 209(4), 645–651. http://doi.org/10.1016/j.amjsurg.2014.12.013

D'Angelo, A. L. D., Rutherford, D. N., Ray, R. D., Mason, A., & Pugh, C. M. (2015). Operative skill: Quantifying surgeon's response to tissue properties. Journal of Surgical Research, 198(2), 294–298. http://doi.org/10.1016/j.jss.2015.04.078

Datta, V., Bann, S., Mandalia, M., & Darzi, A. (2006). The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. American Journal of Surgery, 192(3), 372–378. http://doi.org/10.1016/j.amjsurg.2006.06.001

Davids, K., Button, C., & Bennett, S. (2008). Dynamics of Skill Acquition - A Constraints-Led Approach. Champaign, IL, USA: Human Kinetics.

Deal, S. B., Lendvay, T. S., Haque, M. I., Brand, T., Comstock, B., Warren, J., & Alseidi, A. (2016). Crowd-sourced assessment of technical skills: An opportunity for improvement in the assessment of laparoscopic surgical skills. American Journal of Surgery, 211(2), 398–404. http://doi.org/10.1016/j.amjsurg.2015.09.005

Dobson, R. (2005). The loneliness of the left handed surgeon. BMJ, 330(7481), 10. http://doi.org/10.1136/bmj.330.7481.10-f

Emerson, J. W., & Arnold, T. B. (2011). Statistical sleuthing by leveraging human nature: A study of olympic figure skating. American Statistician, 65(3), 143–148. http://doi.org/10.1198/tast.2011.10165

Emerson, J. W., Seltzer, M., & Lin, D. (2009). Assessing Judging bias: An example from the 2000 Olympic Games. American Statistician, 63(2), 124–131. http://doi.org/10.1198/tast.2009.0026

Evans, J., & Stanovich, K. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. Perspectives on Psychological Science, 8(3), 223–241. http://doi.org/10.1177/1745691612460685

Fitts, P. M., & Posner, M. L. I. (1967). Human Performance. Belmont, CA: Brooks/Cole Publishing Co; Retrieved from http://www.worldcat.org/title/human-performance/oclc/00480476

Frasier, L. L., Azari, D. P., Ma, Y., Quamme, S. R. P., Radwin, R. G., Pugh, C. M., … Greenberg, C. C. (2016). A marker-less technique for measuring kinematics in the operating room. Surgery. http://doi.org/10.1016/j.surg.2016.05.004

French, A., Lendvay, T. S., Sweet, R. M., & Kowalewski, T. M. (2017). Predicting surgical skill from the first N seconds of a task: value over task time using the isogony principle. International Journal of Computer Assisted Radiology and Surgery, 12(7), 1161–1170. http://doi.org/10.1007/s11548-017-1606-5

George, B. C., Teitelbaum, E. N., Meyerson, S. L., Schuller, M. C., Darosa, D. A., Petrusa, E. R., … Fryer, J. P. (2014). Reliability, validity, and feasibility of the zwisch scale for the assessment of intraoperative performance. Journal of Surgical Education, 71(6), e90–e96. http://doi.org/10.1016/j.jsurg.2014.06.018

George, E. I., Skinner, A., Pugh, C. M., & Brand, T. C. (2018). Performance Assessment in Minimally Invasive Surgery. In Surgeons as Educators (pp. 53–91). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-64728-9_5

Ghasemloonia, A., Maddahi, Y., Zareinia, K., Lama, S., Dort, J. C., & Sutherland, G. R. (2017). Surgical Skill Assessment Using Motion Quality and Smoothness. Journal of Surgical Education, 74(2), 295–305. http://doi.org/10.1016/j.jsurg.2016.10.006

Glarner, C. E., Hu, Y. Y., Chen, C. H., Radwin, R. G., Zhao, Q., Craven, M. W., … Greenberg, C. C. (2014). Quantifying technical skills during open operations using video-based motion analysis. Surgery (United States), 156(3), 729–734. http://doi.org/10.1016/j.surg.2014.04.054

Grober, E. D., Roberts, M., Shin, E. J., Mahdi, M., & Bacal, V. (2010). Intraoperative assessment of technical skills on live patients using economy of hand motion: establishing learning

curves of surgical competence. American Journal of Surgery, 199(1), 81–85. http://doi.org/10.1016/j.amjsurg.2009.07.033

Groopman, J. (2008). How Doctors Think. Mariner Books.

Gumbs, A. A., Hogle, N. J., & Fowler, D. L. (2007). Evaluation of Resident Laparoscopic Performance Using Global Operative Assessment of Laparoscopic Skills. Journal of the American College of Surgeons, 204(2), 308–313. http://doi.org/10.1016/j.jamcollsurg.2006.11.010

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutorials in Quantitative Methods for Psychology, 8(1), 23–34. http://doi.org/10.20982/tqmp.08.1.p023

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning. The Mathematical Intelligencer, 27(2), 83–85. http://doi.org/10.1198/jasa.2004.s339

Hatala, R., Cook, D. A., Brydges, R., & Hawkins, R. (2015). Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. Advances in Health Sciences Education, 20(5), 1149–1175. http://doi.org/10.1007/s10459-015-9593-1

ISU. (2017). Communication No. 2098. International Skating Union.

Jelovsek, J. E., Kow, N., & Diwadkar, G. B. (2013). Tools for the direct observation and assessment of psychomotor skills in medical trainees: A systematic review. Medical Education, 47(7), 650–673. http://doi.org/10.1111/medu.12220

Kahneman, D. (2013). Thinking Fast and Slow (1st ed.). New York: Farrar, Strauw and Giroux.

Kane, M. (2006). Validation. In R. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). The Argument-Based Approach to Validation. School Psychology Review, 42(4), 448–457.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine, 15(2), 155–163. http://doi.org/10.1016/j.jcm.2016.02.012

Krippendorff, K., & Krippendorff. (2010). Krippendorff's alpha. In Encyclopedia of Research Design. Thousand Oaks, CA: SAGE Publications, Inc.

Kutner, M. H., Nachtsheim, Christopher, J., Neter, J., & Li, W. (2005). Applied Linear Statistical Models. (B. Gordon, R. T. H. J. Hercher, & L. Stone, Eds.) (5th ed.). McGraw-Hill. Retrieved from http://books.google.fr/books?id=0xqCAAAACAAJ&dq=intitle:Applied+linear+statistical+models+djvu&hl=&cd=1&source=gbs_api

Lacquaniti, F., Terzuolo, C., & Viviani, P. (1983). The law relating the kinematic and figural aspects of drawing movements. Acta Psychologica, 54(1–3), 115–130. http://doi.org/10.1016/0001-6918(83)90027-6

Laufer, S., D'Angelo, A.-L. D., Kwan, C., Ray, R. D., Yudkowsky, R., Boulet, J. R., … Pugh, C. M. (2016). Rescuing the Clinical Breast Examination. Annals of Surgery, XX(X), 1–6. http://doi.org/10.1097/SLA.0000000000002024

Law, K. E., Jenewein, C. G., Gannon, S. J., DiMarco, S. M., Maulson, L. J., Laufer, S., & Pugh, C. M. (2016). Exploring hand coordination as a measure of surgical skill. Journal of Surgical Research, 205(1), 192–197. http://doi.org/10.1016/j.jss.2016.06.038

Lee, J. (2004). Outlier Aversion in Evaluating Performance: Evidence from Figure Skating.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. Annals of Statistics, 44(3), 907–927. http://doi.org/10.1214/15-AOS1371

Legendre, P. (2010). Coefficient of concordance. In Encyclopedia of Research Design ,vol. 1. (Vol. Vol. 1., pp. 164–169). Los Angeles: SAGE Publications, Inc. http://doi.org/10.4135/9781412961288.n55

Mackenzie, C. F., Watts, D., Patel, R., Yang, S., Hagegeorge, G., Hu, P. F., … Tisherman, S. (2016). Sensor-free Computer-Vision hand-motion entropy and video-analysis of technical performance during open surgery on fresh cadavers: report of methodology and analysis. In Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting (pp. 691–695). http://doi.org/10.1177/1541931213601

Madani, A., Vassiliou, M. C., Watanabe, Y., Al-Halabi, B., Al-Rowais, M. S., Deckelbaum, D. L., … Feldman, L. S. (2017). What Are the Principles That Guide Behaviors in the Operating Room? Annals of Surgery, 265(2), 255–267. http://doi.org/10.1097/SLA.0000000000001962

Maier-Hein, L., Vedula, S., Speidel, S., Navab, N., Kikinis, R., Park, A., … Jannin, P. (2017). Surgical Data Science: Enabling Next-Generation Surgery. Retrieved from https://arxiv.org/abs/1701.06482

Martin, J. A., Regehr, G., Reznick, R., MacRae, H., Brown, M., Murnaghan, H., … Brown, M. (1997). Objective structured assessment of technical skill (OSATS) for surgical residents. British Journal of Surgery, 84(2), 273–278. http://doi.org/10.1002/bjs.1800840237

Mieg, H. A. (2009). Two factors of expertise? Excellence and professionalism of environmental experts. High Ability Studies, 20(1), 91–115. http://doi.org/10.1080/13598130902860432

Mohamadipanah, H., Nathwani, J., Peterson, K., Forsyth, K., Maulson, L., DiMarco, S., & Pugh, C. (2018). Shortcut assessment: Can residents' operative performance be determined in the first five minutes of an operative task? Surgery (United States), 0, 1–6. http://doi.org/10.1016/j.surg.2018.02.012

Mohamadipanah, H., Parthiban, C., Law, K., Nathwani, J., Maulson, L., DiMarco, S., & Pugh, C. (2016). Hand smoothness in laparoscopic surgery correlates to psychomotor skills in virtual reality. BSN 2016 - 13th Annual Body Sensor Networks Conference, 242–246. http://doi.org/10.1109/BSN.2016.7516267

Moulton, C. A., Regehr, G., Lingard, L., Merritt, C., & MacRae, H. (2010). Slowing down to stay out of trouble in the operating room: remaining attentive in automaticity. Acad Med, 85(10), 1571–1577. http://doi.org/10.1097/ACM.0b013e3181f073dd

Murphy, J. G., Torsher, L. C., & Dunn, W. F. (2007). Simulation medicine in intensive care and coronary care education. Journal of Critical Care, 22(1), 51–55. http://doi.org/10.1016/j.jcrc.2007.01.003

Neter, J., Wasserman, W., & Kutner, M. (1990). Ch 8 - 9: Building the Regression Model. In R. T. Hercher & E. Shiell (Eds.), Applied Linear Statistical Models (3rd ed., p. 1181). Boston, MA: Richard D. Irwin, INC.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), Cognitive skills and their Acquisition (Vol. 6, pp. 1–55). Hillsdale, NJ: Erlbaum.

Niitsu, H., Hirabayashi, N., Yoshimitsu, M., Mimura, T., Taomoto, J., Sugiyama, Y., … Takiyama, W. (2013). Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. Surgery Today, 43(3), 271–275. http://doi.org/10.1007/s00595-012-0313-7

Olkin, I., Lou, Y., Stokes, L., & Cao, J. (2015). Analyses of Wine-Tasting Data: A Tutorial. Journal of Wine Economics, 10(01), 4–30. http://doi.org/10.1017/jwe.2014.26

Prasad, N. K., Kvasnovsky, C., Wise, E. S., & Kavic, S. M. (2018). The Right Way to Teach Left-Handed Residents: Strategies for Training by Right Handers. Journal of Surgical Education, 75(2), 271–277. http://doi.org/10.1016/j.jsurg.2017.07.004

Pugh, C. M. (2013). Application of national testing standards to simulation-based assessments of clinical palpation skills. Military Medicine, 178(10 Suppl), 55–63. http://doi.org/10.7205/MILMED-D-13-00215

Pugh, C. M. (2014). Getting a sense for the surgical touch [Video File]. Retrieved from https://www.tedmed.com/talks/show?id=292997

Radwin, R. G., Lee, J. D., & Akkas, O. (2017). Driver Movement Patterns Indicate Distraction and Engagement. Human Factors, 59(5), 1–36. http://doi.org/10.1177/0018720817696496

Reznick, R., Regehr, G., MacRae, H., Martin, J., & McCulloch, W. (1997). Testing technical skill via an innovative “bench station” examination. American Journal of Surgery, 173(3), 226–230. http://doi.org/10.1016/S0002-9610(97)89597-9

Sharma, Y., Bettadapura, V., Ploetz, T., Hammerla, N., Mellor, S., McNaney, R., … Essa, I. (2014). Video Based Assessment of OSATS Using Sequential Motion Textures.

Proceedings of M2CAI 2014. Retrieved from http://di.ncl.ac.uk/publicweb//publications/Sharma-et-al-VideoBasedAssessment.pdf

Stratton, S. M., Liu, Y.-T., Hong, S. L., Mayer-Kress, G., & Newell, K. M. (2007). Snoddy (1926) Revisited: Time Scales of Motor Learning. Journal of Motor Behavior, 39(6), 503–515. http://doi.org/10.3200/jmbr.39.6.503-516

Tibshirani, R. (2015). Recent Advances in Post-Selection Statistical Inference.

Vassiliou, M. C., Feldman, L. S., Andrew, C. G., Bergman, S., Leffondré, K., Stanbridge, D., & Fried, G. M. (2005). A global assessment tool for evaluation of intraoperative laparoscopic skills. American Journal of Surgery, 190(1), 107–13. http://doi.org/10.1016/j.amjsurg.2005.04.004

Vedula, S. S., Ishii, M., & Hager, G. D. (2017). Objective Assessment of Surgical Technical Skill and Competency in the Operating Room. Annual Review of Biomedical Engineering, 19(1), 301–325. http://doi.org/10.1146/annurev-bioeng-071516-044435

Vernez, S. L., Huynh, V., Osann, K., Okhunov, Z., Landman, J., & Clayman, R. V. (2017). C-SATS: Assessing Surgical Skills Among Urology Residency Applicants. Journal of Endourology, 31(S1), S-95-S-100. http://doi.org/10.1089/end.2016.0569

Watson, R. A. (2014). Use of a Machine Learning Algorithm to Classify Expertise: Analysis of Hand Motion Patterns During a Simulated Surgical Task. Academic Medicine : Journal of the Association of American Medical Colleges, 89(8), 1–5. http://doi.org/10.1097/ACM.0000000000000316

Weinbergger, S. E., Duffey, F. D., & Cassel, C. K. (2005). “Practice Makes Perfect” . . . Or Does It? Annals of Internal Medicine, 142, 302–303.

White, L. W., Lendvay, T. S., Holst, D., Borbely, Y., Bekele, A., & Wright, A. (2014). Using crowd-assessment to support surgical training in the developing world. Journal of the American College of Surgeons, 219(4), e40. http://doi.org/10.1016/j.jamcollsurg.2014.07.491

Wiegmann, D. A., Eggman, A. A., ElBardissi, A. W., Parker, S. H., & Sundt, T. M. (2010). Improving cardiac surgical care: A work systems approach. Applied Ergonomics, 41(5), 701–712. http://doi.org/10.1016/j.apergo.2009.12.008

Yule, S., Flin, R., Paterson-Brown, S., & Maran, N. (2006). Non-technical skills for surgeons in the operating room: A review of the literature. Surgery, 139(2), 140–149. http://doi.org/10.1016/j.surg.2005.06.017

Yule, S., & Paterson-Brown, S. (2018). Surgeons ’ Non-technical Skills. Surg Clin N Am, 92(2012), 37–50. http://doi.org/10.1016/j.suc.2011.11.004

5. Modeling performance of open surgical cases

5.0 Manuscript Information

This manuscript is intended for submission to *Surgery*.

5.1 Abstract

This study extends automatic and computer-aided assessment of benchtop suturing tasks to the operating room. Prediction models of expert-rated motion economy and fluidity of motion performance (0-10) were previously created from video of 37 clinicians performing common benchtop suturing tasks. Enabled through computer vision of the hands, these models are tested on 47 video clips of expert-rated suturing and tying tasks completed in the operating room. Video comparison of the operating room was contingent on a clear, consistent view of both the surgeon's hands. The relationship between predictive and observed expert ratings for fluidity of motion (slope = 0.82, intercept = 1.77, R^2 = 0.56) performed better than motion economy (slope = 0.73, intercept = 2.04, R^2 = 0.49), although 85% of ratings for both models were within ±2 of the predicted expert response. Models were sensitive to changing hand postures, dropped ligatures, and poor tissue contact while initiating a stitch. In line with Kane and Messick's modern validity framework, these results suggest that performance ratings for suturing and tying tasks extrapolate reasonably well from simulated settings to more complex open surgeries and may helpful to generate formative feedback during deliberate practice on benchtop simulations.

5.2 Background

Objective surgical assessment is increasingly enabled by "surgical data science" (Maier-Hein et al., 2017). Robotic platforms like ROVIMAS (Aggarwal et al., 2007), and the ICSAD (Bann et al., 2003; Datta et al., 2002; Hayter et al., 2009), and virtual reality simulations (Grober et al., 2010), among others, can provide ready fodder for analyzing differences between novice and experienced clinicians. Corvetto et al., (2017), recently found that ICSAD metrics correlated

moderately well with global rating scales of performance on benchtop simulations, and significantly discriminated between experienced and novice performers. Marker-less video tracking hand motion has also demonstrated the ability to discriminate between clinician experience levels in and out of the operating room (Frasier et al., 2016; Glarner et al., 2014; Mackenzie, Watts, Patel, Yang, Garofalo, et al., 2016), and to predict operating room performance along a series of rating scales (Azari et al., 2017).

Limiting comparisons to experienced and novice clinicians, however – the so called "gray hair index" (Cook, 2015) – may be insufficient to uncover what attributes truly comprise surgical skill and performance (Madani et al., 2017). In other words, in comparing expert and novice performance "the absence of hypothesized differences would suggest a serious flaw in the validity argument, but the confirmation of such differences adds little" (Cook, 2015). Modern validity frameworks (Kane, 2006, 2013), instead, require a robust basis of inferences to build a validity argument. David Cook (2015) has provided detailed discussions and guidelines of implementing this validity framework for medical education assessment and examining each inference in turn: scoring (i.e. quantifying observations into scores), generalization (i.e. associating scores and performance in constrained settings), extrapolation (i.e. how scores reflect real-world performance), and implications (i.e. impacts and associated decisions). This approach was adopted by the American Education Research Association (AERA) (Kane, 2006), to support reliable development of objective assessments.

In a broad review of objective computer-aided technical skill evaluation (OCASE-T), Vedula and colleages (2017) discovered a dearth of OCASE-T focused in the operating room, representing a critical gap in extrapolating of assessment scores to meaningful real-world settings. We are in a unique position to leverage marker-less motion tracking of hand movements through video recording – of different settings, and of clinicians with varying experience – to

begin assembling a basis of validity evidence in line with Kane's framework. This study aims to extend previous work predicting surgical performance across various experience categories (scoring and generalization) to the more complex settings of the operating room (extrapolation). These efforts promote a link between observed performance in training situations and observed performance in the operating room – connecting "performances in real life" (Hatala et al., 2015) to those we have automatically quantified in benchtop settings (Azari, 2018).

5.2.1 Assessing Surgery

In 2017, Madani and colleagues, after conducting a cognitive task analysis and broad literature review, rigorously expanded the various domains in which surgeons perform to include 1. psychomotor skills, 2. declarative knowledge, 3. interpersonal skills, 4. personal resourcefulness, and 5. advanced cognitive skills. While the authors provide a much desired structure to address surgical performance, objective assessment and understanding within each of domains is a continuing challenge (Hopmans et al., 2014; Jelovsek et al., 2013; Moorthy, Munz, Sarker, & Darzi, 2003; Williams, Kim, & Dunnington, 2016; Wurzelbacher et al., 2010).

Michael Kane's modern argument-based approach to validity (2006, 2013), accompanied by David Cook's body of work applying Kane's framework in healthcare settings at the Mayo Clinic (Cook et al., 2015, 2014), provides an excellent basis on which to examine the validity of potential assessment measures. Studies of the most widely applied surgical assessment, the Objective Structured Assessment of Technical Skills (OSATS), for instance, have found consistent evidence linking OSATS to a productive use in formative feedback (Hatala et al., 2015). OSATS consists of a series of global rating scales and procedural checklists, and expert assessment of student performance on eight prescribed simulations.

This paper represents a synthesis of work using digital video records of the hands to automatically quantify surgical performance along a series of subjective rating scales (adapted

from the five global rating scales of OSATS). Video from the operating room serves as a testing set for performance along both motion economy and fluidity of motion prediction ratings. We provide a discussion on the accuracy and utility of these models and implications for future training programs.

5.3 Methods

5.3.1 Participants and Settings

Surgeons and students (n = 44) from the University of Wisconsin Hospital and Clinics were recruited to participate through email list-servs, in-person announcements, and recommendation. The University of Wisconsin Institutional Review Board approved each study. Prior to participation, each person completed a survey detailing the following demographic information: years of experience, handedness (i.e. right or left handed), specialty, and amount of training.

Each participant completed a series of suturing and tying tasks, but only those surgeons who agree to be recorded sunder IRB approved protocol were filmed. Video data was collected across two settings, seen in Figure 25 and summarized in Table 10. Of collected video, a total of 5 hours and 58 minutes of video records were selected for motion tracking and subsequent analysis for this study.

5.3.2 Video Processing

We utilized custom software (Chia-Hsiung Chen et al., 2014) to track a region of interest (ROI) in each video frame over the course of a video (Figure 26). We have previously used this technology to observe differences between dominant and non-dominant hands (Glarner et al., 2014), identify differences between tasks and roles (Frasier et al., 2016) and predict expert rated performance during short video clips from the operating room (Azari et al., 2017).

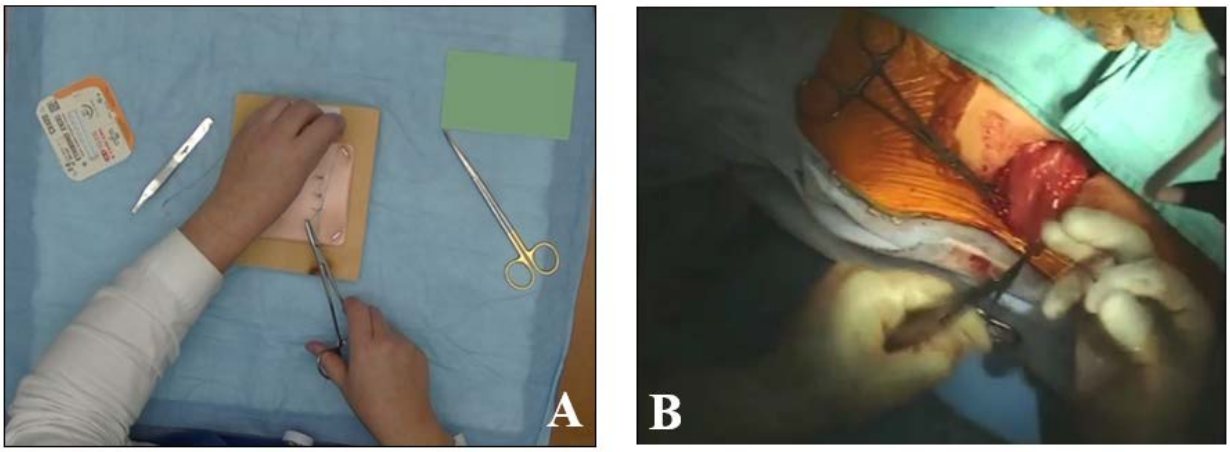


Figure 25: Top down view of common suturing tasks on foam (A) and of operating room (B)

Table 10: Number of participants and length of video recorded by task and material. MS, Medical student; JR, Junior resident; SR, Senior resident; AT, Attending; RT, Retired; SI, Simple interrupted suturing; RS, Running subcuticular suturing; S, General suturing; T, General tying.

Material	Task	MS	JR	SR	JA	SA	RT	Total	Video (hh:mm:ss)
Foam (A)	SI, RS	10	5	10	4	6	2	37	05:47:58
Operating room (B)	S, T	0	0	2	2	3	0	7	00:10:19
Totals (A+B)		10	5	12	6	9	2	44	05:58:17

Figure 26: Region of interest (ROI) to track motion of participant's non-dominant hand while operating.

Each video clip was calibrated to real-world distances using a notecard, ruler, or suture packet of known size. In the OR, the proximal interphalangeal joint breadths for males and females (Figure 27) provided calibration when no object of known size was visible. Pixel locations were identified through Multimedia Video Task Analysis (MVTA) software specially developed at the University of Wisconsin-Madison (Yen & Radwin, 2007). We have used hand-breadth (Akkas et al., 2014) and joint breadth (Frasier et al., 2016) measurements with low coefficients of variation (Greiner, 1991) as calibration coefficients. MVTA allowed us to measure the objects in the video and calibrate pixel measurements to real-world units.

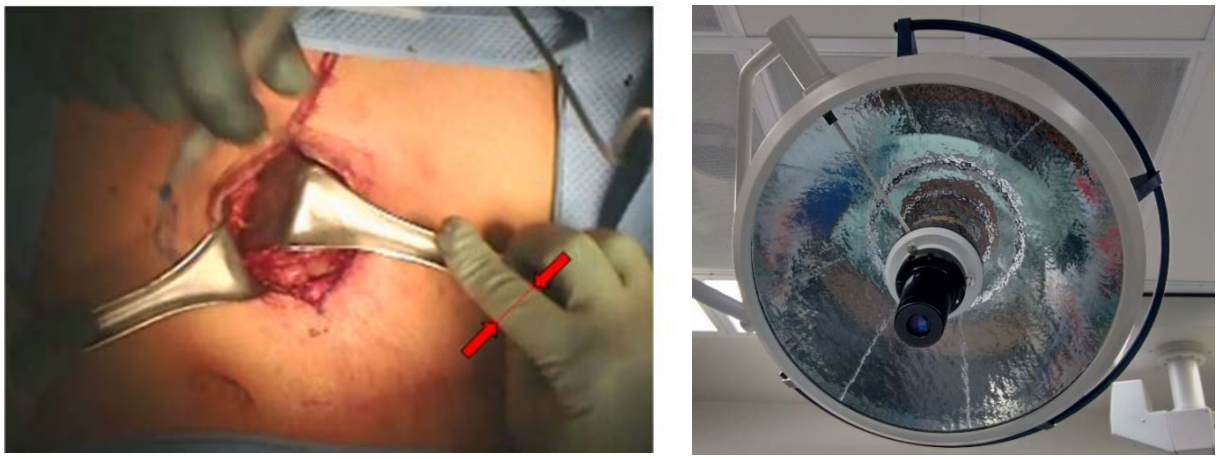


Figure 27: Pixel to real-world calibration (top-left) using PIJB (Greiner, 1991) (bottom-left) if no standard markers are visible from operating room light-mounted camera system (right).

5.3.3 Expert Rated Performance

Video records of participants in Table 10 produced forty-eight clips of suturing and tying tasks (mean length = 13.06s) that met the following criteria: (1) clear view of both hands for entire clip, and (2) observable movement among both hands. While limiting, these criteria are necessary to allow prediction models a similar basis of kinematic data in both settings. Prediction models were drawn directly from those described by Azari (2018).

5.4 Results

For fluidity of motion predictions (Figure 28, slope = 0.83, intercept = 1.75, R^2 = 0.55). The majority of residuals (86%) fell within ±2 of the linear relationship. Motion economy predictions (Figure 29, slope = 0.73, intercept = 2.04, R^2 = 0.49). For this model, 85% of residuals fell within ±2 of the linear relationship. The mean squared errors were 1.58 and 1.74, respectively.

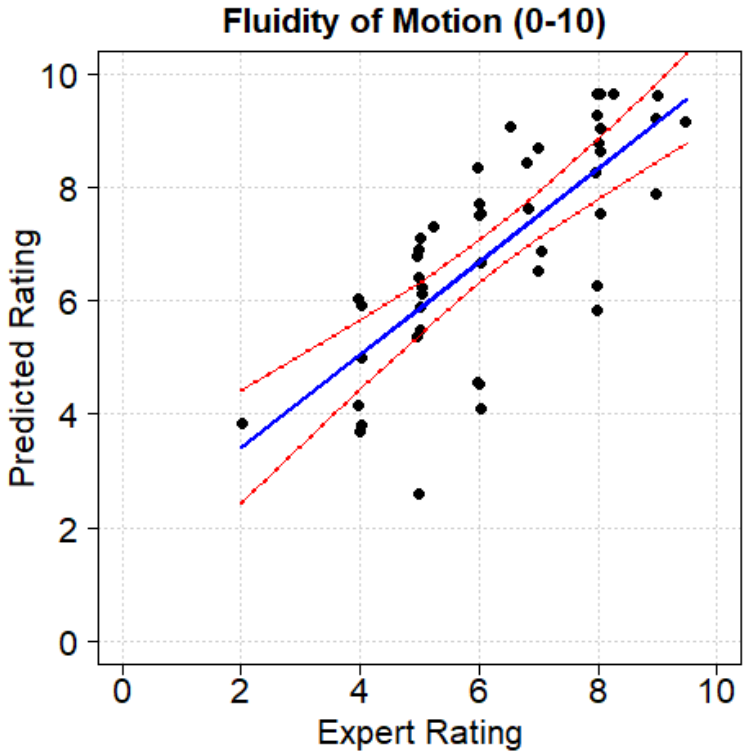


Figure 28: Predicted ratings vs expert ratings for fluidity of motion (0-10) rating scale using (n=48) video clips from the operating room. Confidence intervals (95%) are shown on either side of the linear fit.

Inspection of Figure 28 highlights one substantial underprediction at x = 5, y = 2.5. In the outlier clip, the surgeon changes posture twice, reaching across the patient to gain a better angle of access to suture on bowel tissue. The surgeon changes hand posture prior to actually driving the needle. These contextual adaptations represent an irregular behavior in the training set. On foam, this change could indicate a lack of automaticity or underdeveloped mental-schema to sufficiently represent the task. In the operating room, however, where access to the body is constrained, deliberate or purposeful postural changes in advance of suturing on sensitive tissue

could indicate familiarity with the task constraints. Repeated, indecisive changes in posture or balance in the OR, meanwhile, could indicate a lack of confidence or preparedness.

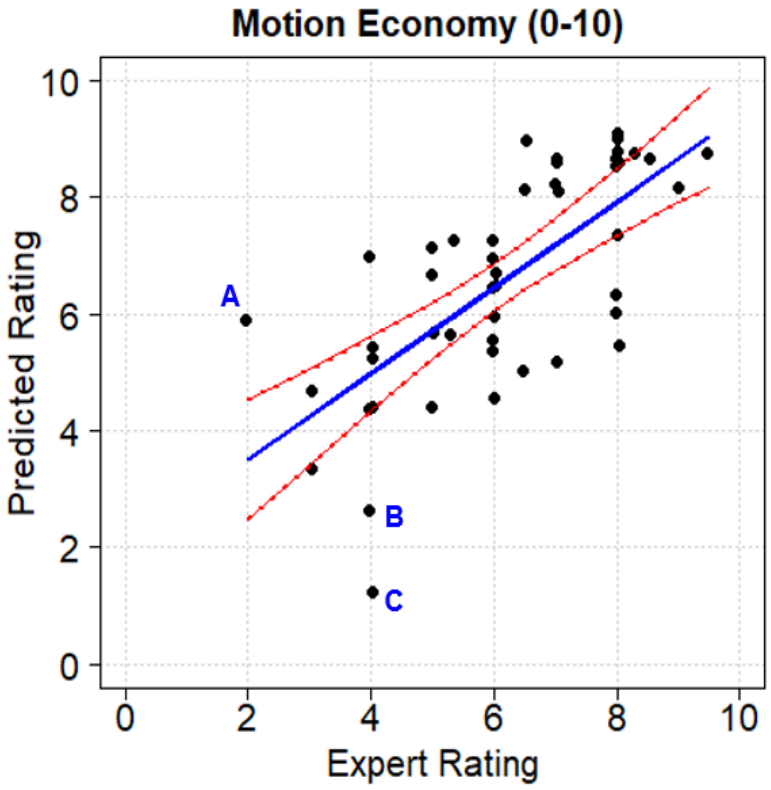


Figure 29: Predicted ratings vs expert ratings for motion economy (0-10) rating scale using (n=48) video clips from the operating room. Confidence intervals (95%) are shown on either side of the linear fit.

Motion economy predictions exhibit greater deviation from the linear relationship between expert and predicted ratings (R^2 = 0.47). Notable outliers in the motion economy predictions (A) x = 2, y = 5.9, (B) x = 4, y = 2.6, and (C) x = 4, y = 1.2, exhibited the following characteristics: (A) multiple missed suturing attempts while driving a needle, (B) a dropped ligature and (C) changing position and posture to access tissue.

Neither prediction performed as well as their benchtop counterparts (slope = 0.71, intercept = 1.98, R^2 = 0.77, correlation = 0.88, R^2_{pred} = 0.70 for fluidity of motion, and slope = 0.65, intercept = 2.36, R^2 = 0.66, correlation = 0.81, R^2_{pred} = 0.61 for motion economy).

5.5 Discussion

This study has shown that automatically generated performance ratings in benchtop settings have real-world relevance in the operating room, albeit with greater variance and sensitivity to contextual factors like missing the tissue while driving a needle, changing postures to fit the hands inside the body cavity, and dropped ligatures. Fluidity of motion predictions performed better than motion economy predictions, and while the majority of residuals fell within a range of ±2, both bench-top derived prediction models exhibited relatively low R^2 values on the real-world data.

Testing each model on video from the OR was hindered by the limited field of view of the overhead light. Forty-seven clips of active suturing and tying in the operating room had a clear and consistent view of both the surgeon's hands. While these clips provide an excellent opportunity to test how well prediction models of benchtop tasks extrapolate to the OR, they do not account for all variations of activity in the OR, nor for all experience or skill levels, or overall case complexity. Future advances in OR video recording, however, may reduce this constraint and build out a bank of video-based surgical motion patterns.

Examination of the outliers for both predictions suggests that the models make no distinction between errors and intentional movements, and how they may be perceived differently in context. A participant suturing on foam, for instance, may be more likely to drive a needle too deeply into the foam as opposed to glancing only the surface and having to reset the needle driver, where as an experienced surgeon in the operating room may prefer to reset the needle driver, rather than penetrate a sensitive tissue too deeply. These differences have significant implications for live surgery, and minimal implications for foam. They are not included in the automatic tracking routine and produce outliers in the new prediction. The motion tracking algorithm also makes no distinction between maneuver technique, such as

handed or tool-based tying. Different tool use may also be contributing to the reduced precision in the new setting.

Segmenting video by the tool used or the type of tissue, as we have explored in the past (Frasier et al., 2016), may improve the relationship between predicted and expert ratings for two-handed performance models in the operating room. However, this increases the burden on manual labeling of video data and limits the potential for comparison between short clips observed in the operating room and video records collected during benchtop trials. As a result, this study does not differentiate between suturing and tying maneuvers within the simulated suturing task. We apply all models equally in the operating room regardless of maneuver or tissue. Parallel work described by Azari (2018) has automatically classified suturing and tying periods every two seconds with 79% accuracy, and may be utilized in the future to further refine measures of fluidity of motion and motion economy within the simulated setting to address this limitation. Future work may also explore how different simulated tissues (e.g. foam, balloons, tissue paper) affect fluidity of motion and motion economy prediction models.

These results of this study emphasize two things: first, that the performance predictions do resemble, on average, real-world operations in a more complex task environment, and second, that they are sensitive to context within the more complex environment, producing outliers under conditions of significant changes in posture or multiple attempts. The different approach towards achieving consensus in previous OR ratings (Azari et al., 2017) and removing outliers (Azari, 2018) also reduces consistency between the expert-generated and predicted ratings. In the terms of Kane and Messick's modern validity framework, these performance scores extrapolate moderately well to the meaningful situations of the operating room but should be limited to providing formative feedback in training scenarios. A "reliability index" could also be included during a formative exercise, to express the relative confidence of the prediction model to the

participant in training. Coaches and trainers may benefit from working with students who have already practiced suturing and tying with the support of automated feedback about how their fluidity of motion and motion economy may be perceived in a real-world operating room. It may also be possible to reverse-engineer these kinds of scores to examine what motion properties are most salient to raters or coaches.

5.6 Conclusion

This study extrapolates benchtop performance prediction models of fluidity of motion and motion economy to the more variable and complex real-world operating room. Prediction models are derived from expert ratings of video clips in each setting. Results are framed within Kane and Messick's modern validity framework and suggest that computer vision of the hands during common benchtop suturing tasks could provide automatic, quantitative feedback of medical student and resident suturing performance. The prediction models provide a reasonable estimation of an average expert rating in the operating room, but do not account for contextual factors, or identify errors.

5.7 References

Aggarwal, R., Grantcharov, T., Moorthy, K., Milland, T., Papasavas, P., Dosis, A., … Darzi, A. (2007). An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. *Annals of Surgery*, *245*(6), 992–999. http://doi.org/10.1097/01.sla.0000262780.17950.e5

Akkas, O., Azari, D. P., Chen, C.-H. E. H. E., Hu, Y. H., Ulin, S. S., Armstrong, T. J., … Radwin, R. G. (2014). A hand speed – duty cycle equation for estimating the ACGIH hand activity level rating. *Ergonomics*, *58*(2), 184–194. http://doi.org/10.1080/00140139.2014.966155

Azari, D. P. (2018). *Quantifying Surgical Skill*. University of Wisconsin-Madison.

Azari, D. P., Frasier, L. L., Quamme, S. R. P., Greenberg, C. C., Pugh, C. M., Greenberg, J. A., & Radwin, R. G. (2017). Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. *Annals of Surgery*, *XX*(Xx), 1. http://doi.org/10.1097/SLA.0000000000002478

Bann, S. D., Khan, M. S., & Darzi, A. W. (2003). Measurement of surgical dexterity using motion analysis of simple bench tasks. *World Journal of Surgery*, *27*(4), 390–394.

http://doi.org/10.1007/s00268-002-6769-7

Chen, C.-H., Hu, Y. H., & Radwin, R. G. (2014). A motion tracking system for hand activity assessment. In *2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)* (pp. 320–324). IEEE. http://doi.org/10.1109/ChinaSIP.2014.6889256

Cook, D. A. (2015). Much ado about differences: why expert-novice comparisons add little to the validity argument. *Advances in Health Sciences Education : Theory and Practice*, *20*(3), 829–34. http://doi.org/10.1007/s10459-014-9551-3

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575. http://doi.org/10.1111/medu.12678

Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education*, *19*(2), 233–250. http://doi.org/10.1007/s10459-013-9458-4

Corvetto, M. A., Fuentes, C., Araneda, A., Achurra, P., Miranda, P., Viviani, P., & Altermatt, F. R. (2017). Validation of the imperial college surgical assessment device for spinal anesthesia. *BMC Anesthesiology*, *17*(1), 1–7. http://doi.org/10.1186/s12871-017-0422-3

Datta, V., Chang, A., Mackay, S., & Darzi, A. (2002). The relationship between motion analysis and surgical technical assessments. *The American Journal of Surgery*, *184*(1), 70–73. http://doi.org/10.1016/S0002-9610(02)00891-7

Frasier, L. L., Azari, D. P., Ma, Y., Quamme, S. R. P., Radwin, R. G., Pugh, C. M., … Greenberg, C. C. (2016). A marker-less technique for measuring kinematics in the operating room. *Surgery*. http://doi.org/10.1016/j.surg.2016.05.004

Glarner, C. E., Hu, Y. Y., Chen, C. H., Radwin, R. G., Zhao, Q., Craven, M. W., … Greenberg, C. C. (2014). Quantifying technical skills during open operations using video-based motion analysis. *Surgery (United States)*, *156*(3), 729–734. http://doi.org/10.1016/j.surg.2014.04.054

Greiner, T. M. (1991). Hand Anthropometry of U.S. Army Personell. *Technical Report Natick, TR-92/011*, 434.

Grober, E. D., Roberts, M., Shin, E. J., Mahdi, M., & Bacal, V. (2010). Intraoperative assessment of technical skills on live patients using economy of hand motion: establishing learning curves of surgical competence. *American Journal of Surgery*, *199*(1), 81–85. http://doi.org/10.1016/j.amjsurg.2009.07.033

Hatala, R., Cook, D. A., Brydges, R., & Hawkins, R. (2015). Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. *Advances in Health Sciences Education*, *20*(5), 1149–1175. http://doi.org/10.1007/s10459-015-9593-1

Hayter, M. A., Friedman, Z., Bould, M. D., Hanlon, J. G., Katznelson, R., Borges, B., & Naik, V. N. (2009). Validation of the Imperial College Surgical Assessment Device (ICSAD) for labour epidural placement. *Canadian Journal of Anesthesia*, *56*(6), 419–426.

http://doi.org/10.1007/s12630-009-9090-1

Hopmans, C. J., Den Hoed, P. T., Van Der Laan, L., Van Der Harst, E., Van Der Elst, M., Mannaerts, G. H. H., … Ijzermans, J. N. M. (2014). Assessment of surgery residents' operative skills in the operating theater using a modified Objective Structured Assessment of Technical Skills (OSATS): A prospective multicenter study. *Surgery (United States)*, *156*(5), 1078–1088. http://doi.org/10.1016/j.surg.2014.04.052

Jelovsek, J. E., Kow, N., & Diwadkar, G. B. (2013). Tools for the direct observation and assessment of psychomotor skills in medical trainees: A systematic review. *Medical Education*, *47*(7), 650–673. http://doi.org/10.1111/medu.12220

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). The Argument-Based Approach to Validation. *School Psychology Review*, *42*(4), 448–457.

Mackenzie, C. F., Watts, D., Patel, R. R., Yang, S., Garofalo, E., Puche, A. C., … Tisherman, S. A. (2016). Sensor-Free Computer Vision Hand-Motion Entropy and Video Analysis of Technical Performance During Open Vascular Surgery: Proof of Concept for Methodology. *Journal of the American College of Surgeons*, *223*(4), e63. http://doi.org/10.1016/j.jamcollsurg.2016.08.166

Madani, A., Vassiliou, M. C., Watanabe, Y., Al-Halabi, B., Al-Rowais, M. S., Deckelbaum, D. L., … Feldman, L. S. (2017). What Are the Principles That Guide Behaviors in the Operating Room? *Annals of Surgery*, *265*(2), 255–267. http://doi.org/10.1097/SLA.0000000000001962

Maier-Hein, L., Vedula, S., Speidel, S., Navab, N., Kikinis, R., Park, A., … Jannin, P. (2017). Surgical Data Science: Enabling Next-Generation Surgery. Retrieved from https://arxiv.org/abs/1701.06482

Moorthy, K., Munz, Y., Sarker, S. K., & Darzi, A. (2003). Objective assessment of technical skills in surgery. *British Medical Journal (BMJ)*, *327*(7422), 1032–1037. http://doi.org/10.1136/bmj.327.7422.1032

Vedula, S. S., Ishii, M., & Hager, G. D. (2017). Objective Assessment of Surgical Technical Skill and Competency in the Operating Room. *Annual Review of Biomedical Engineering*, *19*(1), 301–325. http://doi.org/10.1146/annurev-bioeng-071516-044435

Williams, R. G., Kim, M. J., & Dunnington, G. L. (2016). Practice Guidelines for Operative Performance Assessments. *Annals of Surgery*, *XX*(X), 1. http://doi.org/10.1097/SLA.0000000000001685

Wurzelbacher, S., Burt, S., Crombie, K., Ramsey, J., Luo, L., Allee, S., & Jin, Y. (2010). A comparison of assessment methods of hand activity and force for use in calculating the ACGIH(R) hand activity level (HAL) TLV(R). *Journal of Occupational and Environmental Hygiene*, *7*(7), 407–416. http://doi.org/10.1080/15459624.2010.481171

Summary

This section presents a discussion of the results from chapters 1 through 5. The first portion discusses changes in observed motion features as clinicians gain years of experience. The second section describes progress in automatically predicting surgical maneuvers from continuous video and suggests directions to improve future surgical video collection and processing. The third discusses the success automatically predicting expert-rated performance for both benchtop and operating room procedures. Lastly is a discussion of the implications of the combined body of work in line with the nascent model of quantified performance introduced in the first chapter to support objective and computer aided skill evaluation. I also provide an overview of conceptual software design attributes and potential military implications of this work.

Hand motion changes with experience

Across the three task settings (foam, pig feet, bowel), experience was associated with several changes in hand motion. As found in past studies, these were not reflected evenly among dominant and non-dominant hands (Glarner et al., 2014). Medical students and residents often exhibited differences in speed and acceleration for dominant hand use, while attendings and residents exhibited different path length, speed, and acceleration in their non-dominant hands. Residents, for example, exhibited significantly less path length per cycle of both hands and greater speeds for the dominant hand, resulting in increased cycle frequency of completion than medical students. Attending surgeons exhibited a similar increase in cycle frequency compared to junior residents, and reductions in path length per cycle of the non-dominant hand over residents in general. Attending surgeons also progressed through all stages of the task more quickly, while accelerating less frequently, accelerating less overall, and generally moving slower while tying with their non-dominant hands than both residents and medical students. Both

path length and cycle frequency exhibited clear trends across all seven experience categories, despite significant differences occurring only between junior status of one role and senior status of the following role. These trends suggest that future studies may be able to apply categorical linear regression models to establish validity evidence of generalization across tasks, rather than segmenting the population into discrete buckets as traditionally done.

Machine learning classifies surgical maneuvers

Even though video is easy to collect, analyzing video has many challenges. Health care settings easily capture more video than can be reviewed and may not be equipped to address more technical aspects of processing. Video formats, frame rates, compression schemes, and compatibility issues pose significant challenges to on-demand review, consistent time indexing, and frame-by-frame motion analysis. Automatic deconstruction of surgical video to expedite video review would help offset these challenges. This study trains machine learning models to classify the surgical maneuvers of suturing and tying, and the transitions between those maneuvers, with 79% accuracy of all two-second periods using a combination of random forest and hidden markov models on a reserved subset of participants. This is consistent with other three-state maneuver models for robotic and laparoscopic procedures and increases the potential of being able to scroll through a pre-labeled video record to select periods of interest for student or coach review. These results were also able to strongly predict the frequency of activity for a set of reserved participants (slope = 0.88, intercept = 0.03, correlation = 0.83, R^2 = 0.72), intersecting well the studies utilizing cycle frequency in categorizing experience and predicting performance. Additional techniques, such as conditional random fields (CRF) are introduced in the study, which may yield improvements for future video analysis of longer or more complex tasks collected in the operating room.

Predicting performance in and out of the operating room

Chapters 4 and 5 study how predicted expert-ratings of four visual analog scales (0-10) of fluidity of motion, motion economy, tissue handling, and hand coordination generalize across participants on benchtop simulations and extrapolate to real-world settings. These scales were designed to encompass the full range of observable behavior during short (< 90 s) clips of surgical maneuvers. Fluidity of motion exhibited the best relationship between predicted and expert ratings (slope = 0.71, intercept = 1.98, R^2 = 0.77, correlation = 0.88, R^2_{pred} = 0.70), and also provided the most consistent extrapolation to real-world operating room settings (slope = 0.83, intercept = 1.75, R^2 = 0.55), albeit with greater variation as seen by the reduced R^2 value. Motion economy predictions in benchtop settings (slope = 0.65, intercept = 2.36, R^2 = 0.66, correlation = 0.81, R^2_{pred} = 0.61) extrapolated moderately well to the operating room (slope = 0.73, intercept = 2.04, R^2 = 0.49), but were more sensitive to postural changes within the constrained working area. Tissue handling and hand coordination scores appeared most sensitive to contextual factors and were not tested in the operating room setting. These results assemble generalization and extrapolation evidence of validity for fluidity of motion and motion economy in line with Kane and Messick's modern framework. They provide a connection between those benchtop simulation performance and the real-world operating room.

Implications for surgical training

Real time capture of surgical motion on video represents great opportunity for artificial intelligence and machine learning algorithms to automatically segment, process, and assess performance before patients are put at risk. Chapter 1 provides a consistent lexicon with which to describe performance gains across a surgeon's career. Chapter 2 explores observed differences in hand motion across clinicians of varying experience, while chapter 3 advances abilities to automatically classify surgical video into discrete segments. Chapters 4 and 5 provide generalization and extrapolation evidence of automatic and quantified performance models. Each

of these studies addresses a necessary step in building automatic video-based formative assessment tools and promote subsequent development of video-based assessment platforms. Once observed and quantified, properties of surgical hand motion could be fed back to the surgeon through visual or tactile interfaces through a dynamic, on-demand dashboard to promote performance improvement of hands-on clinical skills.

Focusing on performance

Traditional domains of performance – music, art, and athletics, for example – showcase impressive and skills rehearsed and practiced over years of training. They are competitive, employ traditional apprenticeship relationships, and are appreciated subjectively. Surgery shares many such facets: it has been described as a craft (Reznick, 1993; Thomas, 2006), an art (Dartmouth, 2002; Khan, Bann, Darzi, & Butler, 2002), grounded through apprenticeship (Sealy, 1999), and promoting structured social hierarchies favoring years of practice (Bosk, 1979; Williamson, 2004). Historically, the "operating theater" has even reflected a particular form of showmanship and wonderment (Fitzharris, 2017; Frumovitz, 2002).

Just like these other fields, surgeons too, do not always perform at their best. They are subject to biases in judgement and recall (Williams, Klamen, & Mcgaghie, 2003), commit and need to manage errors of various types (Law, 2016), and are influenced by factors including those: (1) outside the control of the surgeon; (2) related directly to the surgeon; and (3) inherent in the particular decision to be made (Francis, 2009). Physical positioning and ergonomics play a role (Kruse, Luebbers, Grätz, & Obwegeser, 2010; Rosenblatt, McKinney, & Adams, 2013; Steinhilber et al., 2015), as do fatigue, caffeine, stress (Fargen, Turner, & Spiotta, 2016), and social relationships with patients and the care team (Rosen et al., 2010; Salas, Cooke, & Rosen, 2008). Out of the sum of these factors – the available context (Feltovich et al., 1997) and work-system (Carayon et al., 2014; Wiegmann et al., 2010) – emerges a unique, albeit temporary effect

on a surgeon's observable behavior. While professional training may emphasize how to "get the job done" in difficult circumstances, it is also unreasonable to expect a surgeon to perform best and manage errors well while exhausted, distracted, or under duress.

In his book, *How Doctors Think,* Jerome Groopman (2008) further describes how these kinds of transitory factors affect physician performance. He argues that such impacts are amplified through changing technology (p. 148), notions of hero worship (p. 72, 145), aspects of human biology that "can't be predicted" (p. 124) and intense engrained biases (p. 59). Groopman warns that such preconceptions produce dangerous situations in which physicians may "become wedded to distorted conclusions" (p. 27), deleteriously impacting case outcomes and artificially limiting maximal performance than what one could otherwise achieve.

Given the gravity of consequences to life and livelihood, understanding surgical performance may benefit from comparisons to other physically, mentally and morally demanding fields like aviation, astronautics, search and rescue, nuclear command and control, and military operations. Like these, surgeries are high-risk activities localized to a unique time and place, requiring extensive training and readiness to engage in suddenly changing circumstances (O'Neil et al., 2014). Outcomes are dependent on a confluence of difficult-to-predict factors; you can't just "try again" if something goes wrong. Each attempt is unique, and there are often as many, if not more ways things can go wrong, than they can go right. Lessons and techniques from other fields – especially engineering (Rutherford et al., 2015) – are increasingly explored as avenues to improve surgical quality and healthcare overall (Gawande, 2011; Gordon, Mendenhall, & O'Connor, 2013). Even in the age of increasing healthcare simulation capabilities (Stefanidis et al., 2015) developing surgical skills on live patients remains an integral part of clinician growth, even as an attending (Birkmeyer et al., 2013; Carty et al., 2009).

Computer vision applied through digital video, offers a cheap and scalable method to measure performance in and out of the operating room while minimizing patient risk. This dissertation improves the ability of computer vision of surgical video to (1) identify changes in hand motion as clinicians gain experience, (2) automatically predict performance both in and out of the operating room, and (3) advance automatic deconstruction and labeling of benchtop task video such that it can be compared with longer-studied robotic and sensor-equipped simulations. These developments are necessary to promote objective and computer aided evaluation of surgical skills through easily collected video records. A medical student wishing to develop their surgical skill could set up a laptop and a webcam over a piece of foam, perform the simulation, and receive a formative assessment report for every trial. They could automatically compare their overall cycle frequency to that of more experienced clinicians, see how their path length per cycle has reduced over time, and monitor the acceleration peak rate in their dominant hand in an attempt to focus attentively on a smoother their trajectory. Increasing these scores, over time, would suggest they are climbing up the learning curve of the quantified performance model and would be better prepared for the meaningful interactions with attendings and coaches (Huang, Wyles, Chern, Kim, & O'Sullivan, 2016), or to take a summative assessment for a particular procedure.

Software Development

The new capabilities detailed in this dissertation naturally lend themselves to development of a responsive computer interface that can efficiently convey observed attributes of performance directly back to the participant or their coach, either in real-time or for after-action review. This section describes the role of software assisted performance and conceptual design features of such a program, as well as current caveats and challenges.

Software assisted performance

It is largely agreed clinician performance can be enhanced in various ways through computation. Decisions using a combination of model outputs and clinician decision outperforms judgements of just the clinician alone (Aldag, 2012; Dawes, 1979; Dawes, Faust, & Meehl, 1989). Kleinmuntz (1992) highlights specific instances of programs which have helped clinicians to treat meningitis, manage chemotherapy, plan facilities and identify ideal antimicrobial treatments. He uses these case examples to say that computers certainly can think and learn in new environments, but that thinking of a computer as a "surrogate clinician" should be handled skeptically, at best. Arkes and colleagues (1986) consistently demonstrated that such computation can improve clinician decision (although at times with low accuracy). There is also beneficial evidence that decisions made with modeling and data outperform their clinician counterparts with greater reliability tracing (Goldberg, 1970). More recently, IBM has been using the supercomputer Watson to synthesize worldwide clinical results and propose novel treatments for cancer patients – albeit with increasing awareness of its limitations to address variations in cancer type (Ross & Swetlitz, 2017). Still, integration of information throughout medicine (Maier-Hein et al., 2017) represents a great opportunity for technological augmentation to help doctors to "keep up" with current research (Francis, 2009), and continue to test their skills throughout their career. Objective and easily collected measures of cycle frequency, fluidity of motion or motion economy would be a welcome tool to support coaching and development of student psychomotor skills; providing a structured opportunity to reflect and focus attention on specific attributes during surgical rehearsal – critical components of deliberate practice.

Implementation challenges

Implementing any assessment (formative or otherwise), however, risks alienating surgeon independence and challenging physician autonomy. As Charles Bosk describes in his book *Forgive and Remember,* doctors are inherently committed to (and primarily responsible for) providing morally sound judgements: "when things go awry, when the professional's efforts to aid his client fail whatever the reason, the professional's last line of defense –should he doubt himself, should his colleagues question him, should his clients or his representatives accuse him – is that he did everything possible....a moral defense, not a technical one" (p. 164.) Despite Bosk writing in the late 1970's – a time that contributed, in part, to obscuring the gender of the female surgeons he worked with (Williamson, 2004) – he highlights the discrepancy between best practice standards and those derived from medical professionalism that Harald Mieg identified as late as 2009: "As far as control of performance is concerned, we would expect impersonal evaluations of techniques to have priority over personal judgments of individual's moral performance. How are we to account for the fact that the opposite is the case?" (160.)

Even in the low-states arena of formative assessment, implementing assessment techniques needs to be sufficiently framed such that dignity interests are respected. These challenges are plentiful throughout healthcare settings. In his book *Checklist Manifesto,* Atul Gawande (2011) argues that surgical outcomes may be improved rapidly and with low cost, by some of the most useful evidence-based practices guidelines from other professions, especially aviation (i.e. adapted checklists capable of lowering deaths by 1/3, or in some cases, by 47%). However, he builds the case, that data-based improvements and attempts to measure performance require a fundamental change to the culture of medicine, surgery and team roles. In her book, *Beyond the Checklist*, author Suzanne Gordon argues that cultural challenges must not prevent improvements to quality care: "No one can prove who experiences more job stress or complex

responsibility, and in the end, this is a spurious debate…if one industry can benefit from the experience of the other and reduce errors and thus enhance safety, why wouldn't it try?" (Gordon, Mendenhall, & O'Connor, 2013).

The presentation, style, framing, and general interpretation of formative feedback during potential training, as an example of such efforts, represents a crucial vector to ensure progress in clinical medical education (Ende, 1983). If feedback on a surgical assessment is inappropriately timed or configured, they may be more detrimental and unintentionally undermine support for training interventions. As Karsh (2004) describes, the effectiveness of new medical information technology could be limited by general dissatisfaction and ignoring existing workflows (on in a training scenario, expectations) within a work-system approach. Future development of a well-received tool interface may benefit from the design parameters put forth by Brown and Bell in their paper "Authoring Adaptive Tutors for Simulations in Psychomotor Skills Domains" (Brown & Bell, 2017). Considering these implementation challenges and design suggestions will help to ensure that any tool is adopted and used with greater overall utility.

Conceptual design

A hypothetical feedback interface is depicted in Figure 30. This interface would provide relevant metrics that enable (1) continuing analysis of surgical hand motion, and (2) independent learning through directed, deliberate practice suggestions. Current considerations for such an interface are described in turn.

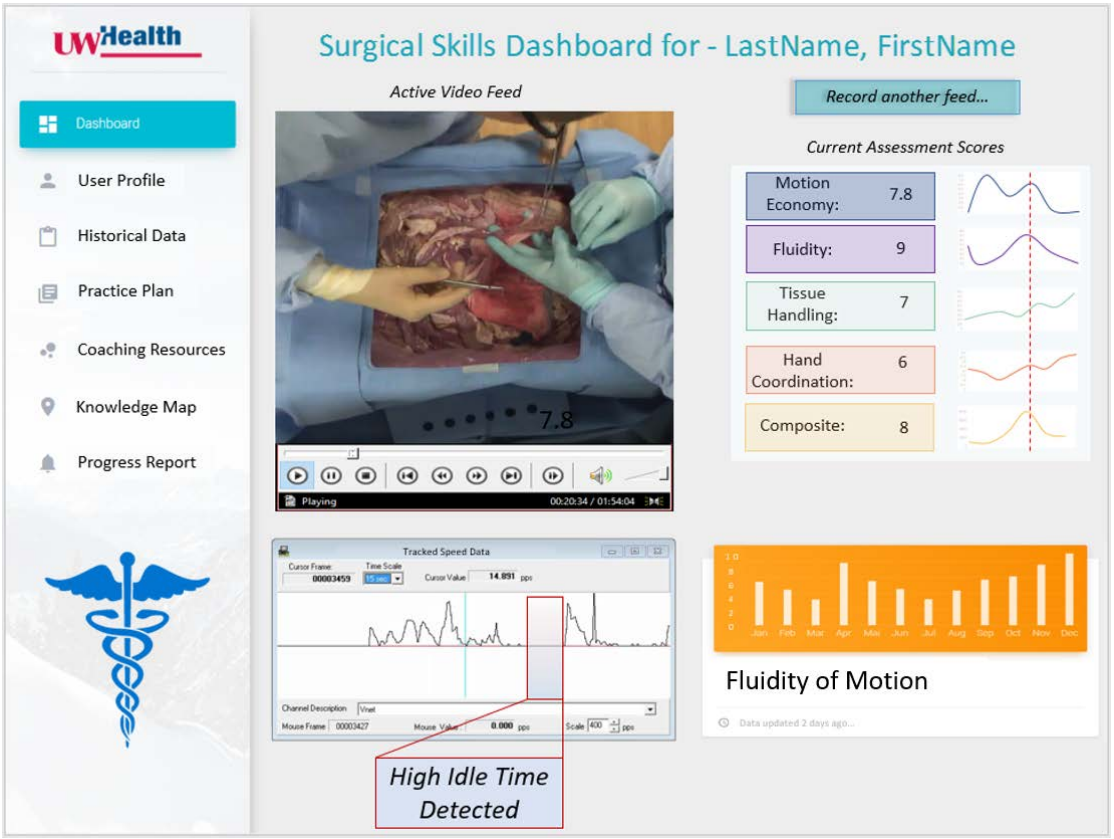


Figure 30: Conceptual dashboard design, for future interactive surgical skills quantification.

(1) On-demand Recording and Tracking - A participant could perform a simulated procedure under a connected webcam, view, and save the digital video of their hands. Motion tracking would be conducted in real or near-real time. The video field would show multiple time-synchronized views (if available). Additional sensors or depth cameras would similarly be time-synchronized.

(2) Multi-Mode Recording – A participant would be able to designate each trial as a practice or assessment round. This packaging would encourage reflection throughout recorded "practice periods" for all skill levels.

(3) On-demand Formative Assessment – Objective assessment scores such as fluidity of motion (0-10) and motion economy (0-10) would be reported, with contributing kinematic features to each score identified accessible via drilldown. Any lagging or

underperforming attribute would be automatically brought to the forefront of the screen and the measure shown, along with the current score.

(4) Practice Plan – Based on current performance, a host of "top suggestions" would automatically populate, accessible under individually customizable "practice plan." These suggestions would take the form of a series of instructional recorded videos, key tips, animations showing different paths, and any other supporting material deemed useful in facilitating independent skills development.

(5) Coaching Resources – Before (and after) completing a task, the participant will have the option to review a series of audio records / videos of highly experienced surgeons sharing tips, tricks, or examples of specific instances in their career. Such "war stories" (Y. Y. Hu et al., 2012) and verbal feedback from more experienced surgeons (Porte, Xeroulis, Reznick, & Dubrowski, 2007) are instrumental in building career and case expectations.

(6) Historical Data – The participant could see aggregate performance data across all samples within a "progress report."

(7) Visualization – Relevant visualizations, including time plots within task (i.e. speed, acceleration), scatter plots of performance (i.e. score on one scale over several practice periods), relative population performance, and any other relevant summary information would be depicted graphically, preferably using an HTML/CSS interface. The C# Xamarin plugin to develop distributable code base across Windows, iOS and Android platforms may provide an advantage. Several visualizations like path length over time or a time plot vs speed could be loaded automatically, with more specific versions accessible on-demand. These visualizations and summaries could be exported to .pdf or .docx files as part of a "kinematic report card." A growing number

of visualization routines for this type of cyclical behavior (Greene et al., 2017) may be established and tested to support the real-time and on-demand assessment feedback to assist physician psycho-motor development in accordance with these suggestions.

(8) Impact – Physicians could use this system to gauge their own awareness, promote reflection on specific attributes (e.g. fluidity of motion) and examine their performance on a repeated basis throughout their career. Future study will need to solidify the recommendations proposed here, and test efficacy of different design parameters or feedback routines to facilitate surgical training for productive implementation of such an interface.

Military Implications

This type of dashboard would be especially useful for military surgeons who face substantial challenges in leveraging their expertise between active duty and civilian practice (or the reserves). Deployed surgeons may be underprepared to operate in combat situations and on combat-oriented injuries that fall outside their experience; such as truncal hemorrhage or skeletal reconstruction from penetrating injuries (Kelly et al., 2008; Tyler et al., 2010). At the same time, clinical and specialized surgical skills degrade while surgeons are deployed; especially for laparoscopic surgical skills (Perez et al., 2013).

Surgeons who face intermittent mobilizations often have highly specialized practices, yet when deployed to a combat theater, are required to perform acute general and trauma surgery or in some cases, assigned to positions with limited operative opportunities. Deployed surgeons reported that deployment increases trauma skills ($p < 0.001$) but decreases the procedural skills required for civilian practice ($p < 0.005$), taking 3 to 6 months on average to return to pre-deployment skill level upon returning to practice (Deering, Rush, Lesperance, & Roth, 2011).

These professionals face the challenge of maintaining the various procedural skills required for both military service and their civilian practice. While deployed, surgeons may have time to devote to simulation to maintain or refresh specialty-specific technical skills but lack the equipment to do so. Current approaches require sophisticated, expensive hardware only found in specialized simulation settings. This research advances a computational model of surgical performance and automatic video processing which would allow for performance assessment during periods away from practice or professional transition. Such a novel, portable system utilizes hardware that is broadly available, even in combat training settings such as digital video capture and cloud-based or local computer processing.

This dissertation advances objective metrics for surgical tasks that have broad applicability to traumatic battlefield injuries. Access to immediate, reproducible kinematic-based feedback described previously, can inform self-assessment, and direct practice of specific surgical maneuvers and overall surgical performance. This may help to provide a venue in which skill development is quantifiably traceable and intentionally achieved before it is needed. The capacity to deconstruct surgical skill can provide a deeper understanding of the kinematics of surgical performance and aid in skill acquisition, even in difficult to access or remote areas.

Future Challenges

Quantitative observation of hand motion through digital video enables objective understanding of common maneuvers such as suturing and tying. In order to ensure that on-demand performance feedback, will be, as Kleinmuntz (1992) advocates for computer aided healthcare settings “a welcome addition to the physician’s clinical armamentarium,” there are three main axes of necessary improvement not addressed by this thesis. First, the automaticity of tracking algorithm needs to be improved, and amount of manual interventions made by an operator would need to be reduced. This hurdle is ubiquitous, and commonly cited as a barrier to

engaging in marker-less motion tracking studies (Ganni, Botden, Chmarra, Goossens, & Jakimowicz, 2018). Second, audio-visual setup and collection would need to be streamlined, with software processing of variable frame rates, calibration, and compatibility issues resolved automatically. Third, and finally, the prediction and state models created by Azari (2018) would need to be re-packaged to operate either via cloud based video upload, or to run within stand-alone programs. Lastly, seeing benefits in improved patient outcomes as a result of this work depends also on parallel and continuing efforts to promote surgical safety, improve training and coaching techniques, and enhance error detection and management strategies.

References

Aldag, R. J. (2012). Distinguished Scholar Invited Essay Behavioral Decision Making: Implications for Leadership and Organizations. Journal of Leadership & Organizational Studies, 19(2), 133–141. http://doi.org/10.1177/1548051812442745

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. Organizational Behavior and Human Decision Processes, 37(1), 93–110. http://doi.org/10.1016/0749-5978(86)90046-4

Azari, D. P. (2018). Quantifying Surgical Skill. University of Wisconsin-Madison.

Birkmeyer, J. D., Finks, J. F., O'Reilly, A., Oerline, M., Carlin, A. M., Nunn, A. R., … Birkmeyer, N. J. O. (2013). Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine, 369(15), 1434–42. http://doi.org/10.1056/NEJMsa1300625

Bosk, C. L. (1979). Forgive and Remember: Managing Medical Failure (2nd ed.). Chicago: University of Chicago Press.

Brown, D., & Bell, B. (2017). Authoring adaptive tutors for simulations in psychomotor skills domains. In MODSIM World 2017 (pp. 1–10). Retrieved from http://www.modsimworld.org/papers/2017/Authoring_adaptive_tutors_for_simulations_in_psychomotor_skills_domains.pdf

Carayon, P., Wetterneck, T. B., Rivera-Rodriguez, A. J., Hundt, A. S., Hoonakker, P., Holden, R., & Gurses, A. P. (2014). Human factors systems approach to healthcare quality and patient safety. Applied Ergonomics, 45(1), 14–25. http://doi.org/10.1016/j.apergo.2013.04.023

Carty, M. J., Chan, R., Huckman, R., Snow, D., & Orgill, D. P. (2009). A detailed analysis of the reduction mammaplasty learning curve: A statistical process model for approaching

surgical performance improvement. Plastic and Reconstructive Surgery, 124(3), 706–714. http://doi.org/10.1097/PRS.0b013e3181b17a13

Dartmouth. (2002). The Art of Surgery. Dartmouth Medicine, Fall, 29–39.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. American Psychologist, 34(7), 571–582. http://doi.org/10.1037/0003-066X.34.7.571

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. Science, 243(4899), 1668–1674. http://doi.org/10.1126/science.2648573

Deering, S. H., Rush, R. M., Lesperance, R. N., & Roth, B. J. (2011). Perceived effects of deployments on surgeon and physician skills in the US Army Medical Department. The American Journal of Surgery, 201(5), 666–672. http://doi.org/10.1016/j.amjsurg.2011.01.006

Ende, J. (1983). Feedback in Clinical Medical Education. JAMA: The Journal of the American Medical Association. http://doi.org/10.1001/jama.1983.03340060055026

Fargen, K. M., Turner, R. D., & Spiotta, A. M. (2016). Factors That Affect Physiologic Tremor and Dexterity during Surgery: A Primer for Neurosurgeons. World Neurosurgery. http://doi.org/10.1016/j.wneu.2015.10.098

Feltovich, P. J., Ford, K. M., & Hoffman, R. R. (1997). Exertise in Context. Cambridge: The MIT Press.

Fitzharris, L. (2017, October). How Ether Transformed Surgery from a Race against the Clock. Scientific American. Retrieved from https://www.scientificamerican.com/article/how-ether-transformed-surgery-from-a-race-against-the-clock/

Francis, D. M. A. (2009). Surgical decision making. ANZ Journal of Surgery, 79(12), 886–891. http://doi.org/10.1111/j.1445-2197.2009.05139.x

Frumovitz, M. M. (2002). Thomas Eakins' Agnew Clinic: A study of medicine through art. Obstetrics and Gynecology, 100(6), 1296–1300. http://doi.org/10.1016/S0029-7844(02)02368-2

Ganni, S., Botden, S. M. B. I., Chmarra, M., Goossens, R. H. M., & Jakimowicz, J. J. (2018). A software-based tool for video motion tracking in the surgical skills assessment landscape. Surgical Endoscopy and Other Interventional Techniques, 32(6), 2994–2999. http://doi.org/10.1007/s00464-018-6023-5

Gawande, A. A. (2011). The Checklist Manifesto (1st ed.). New York: Metropolitan Books, Henry Holt and Company.

Glarner, C. E., Hu, Y. Y., Chen, C. H., Radwin, R. G., Zhao, Q., Craven, M. W., … Greenberg, C. C. (2014). Quantifying technical skills during open operations using video-based motion analysis. Surgery (United States), 156(3), 729–734. http://doi.org/10.1016/j.surg.2014.04.054

Goldberg, L. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. Psychological Bulletin, 73(6), 422–432. http://doi.org/10.1037/h0029230

Gordon, S., Mendenhall, P., & O'Connor, B. B. (2013). Beyond the Checklist (1st ed.). Ithaca: Cornell University Press.

Greene, R. L., Azari, D. P., Hu, Y. H., & Radwin, R. G. (2017). Visualizing stressful aspects of repetitive motion tasks and opportunities for ergonomic improvements using computer vision. Applied Ergonomics, 65, 461–472. http://doi.org/10.1016/j.apergo.2017.02.020

Groopman, J. (2008). How Doctors Think. Mariner Books.

Hu, Y. Y., Peyre, S. E., Arriaga, A. F., Roth, E. M., Corso, K. A., & Greenberg, C. C. (2012). War stories: A qualitative analysis of narrative teaching strategies in the operating room. American Journal of Surgery, 203(1), 63–68. http://doi.org/10.1016/j.amjsurg.2011.08.005

Huang, E., Wyles, S. M., Chern, H., Kim, E., & O'Sullivan, P. (2016). From novice to master surgeon: improving feedback with a descriptive approach to??intraoperative assessment. American Journal of Surgery, 212(1), 180–187. http://doi.org/10.1016/j.amjsurg.2015.04.026

Karsh, B.-T. (2004). Beyond usability: designing effective technology implementation systems to promote patient safety. Quality & Safety in Health Care, 13(5), 388–94. http://doi.org/10.1136/qhc.13.5.388

Kelly, J. F., Ritenour, A. E., McLaughlin, D. F., Bagg, K. a, Apodaca, A. N., Mallak, C. T., … Holcomb, J. B. (2008). Injury Severity and Causes of Death From Operation Iraqi Freedom and Operation Enduring Freedom: 2003–2004 Versus 2006. Journal of Trauma and Acute Care Surgery, 64(2), S21-S26; discussion S26-S27. http://doi.org/10.1097/TA.0b013e318160b9fb

Khan, M. S., Bann, S. D., Darzi, A., & Butler, P. E. M. (2002). Suturing: A lost art. Annals of the Royal College of Surgeons of England, 84(4), 278–279. http://doi.org/10.1308/003588402320439748

Kleinmuntz, B. (1992). Computers as clinicians: An update. Computers in Biology and Medicine, 22(4), 227–237. http://doi.org/10.1016/0010-4825(92)90062-R

Kruse, A. L. D., Luebbers, H. T., Grätz, K. W., & Obwegeser, J. a. (2010). Factors influencing survival of free-flap in reconstruction for cancer of the head and neck: a literature review. Microsurgery, 30(3), 242–248. http://doi.org/10.1002/micr

Law, K. E. (2016). Intra-operative errors and error management in chief surgical residents: Mechanisms of mistakes and strategies for recovery.

Maier-Hein, L., Vedula, S., Speidel, S., Navab, N., Kikinis, R., Park, A., … Jannin, P. (2017). Surgical Data Science: Enabling Next-Generation Surgery. Retrieved from https://arxiv.org/abs/1701.06482

O'Neil, H. F., Perez, R. S., & Baker, E. L. (2014). Teaching and Measuring Cognitive Readiness. (H. F. O'Neil, R. S. Perez, & E. L. Baker, Eds.)Teaching and Measuring Cognitive Readiness (Vol. 9781461475). Boston, MA: Springer US. http://doi.org/10.1007/978-1-4614-7579-8

Perez, R. S., Skinner, A., Weyhrauch, P., Niehaus, J., Lathan, C., Schwaitzberg, S. D., & Cao, C. G. L. (2013). Prevention of surgical skill decay. Military Medicine, 178(10 Suppl), 76–86. http://doi.org/10.7205/MILMED-D-13-00216

Porte, M. C., Xeroulis, G., Reznick, R. K., & Dubrowski, A. (2007). Verbal feedback from an expert is more effective than self-accessed feedback about motion efficiency in learning new surgical skills. American Journal of Surgery, 193(1), 105–110. http://doi.org/10.1016/j.amjsurg.2006.03.016

Reznick, R. K. (1993). Teaching and testing technical skills. The American Journal of Surgery, 165(3), 358–361. http://doi.org/10.1016/S0002-9610(05)80843-8

Rosen, M. A., Lazzara, E. H., Lyons, R., Salas, E., Mc Keever R. N., M., O., L. A. D., & N., M. B. R. (2010). Does Teamwork Improve Performance in the Operating Room? A Multilevel Evaluation. Number, 36(3), 133–142. Retrieved from All Papers/R/Rosen et al. 2010 - Does Teamwork Improve Performance in the Operating Room - A Multilevel Evaluation.pdf

Rosenblatt, P. L., McKinney, J., & Adams, S. R. (2013). Ergonomics in the operating room: Protecting the surgeon. Journal of Minimally Invasive Gynecology, 20(6), 744. http://doi.org/10.1016/j.jmig.2013.07.006

Ross, C., & Swetlitz, I. (2017). IBM pitched Watson as a revolution in cancer care. It's nowhere close. Retrieved July 28, 2018, from https://www.statnews.com/2017/09/05/watson-ibm-cancer/

Rutherford, D. N., D'Angelo, A.-L. D., Law, K. E., & Pugh, C. M. (2015). Advanced Engineering Technology for Measuring Performance. Surgical Clinics of North America, 95(4), 813–826. http://doi.org/10.1016/j.suc.2015.04.005

Salas, E., Cooke, N. J., & Rosen, M. a. (2008). On teams, teamwork, and team performance: discoveries and developments. Human Factors, 50(3), 540–7. http://doi.org/10.1518/001872008X288457.

Sealy, W. C. (1999). Halsted is dead: Time for change in graduate surgical education. Current Surgery, 56(1–2), 34–39. http://doi.org/http://dx.doi.org/10.1016/S0149-7944(99)00005-7

Stefanidis, D., Sevdalis, N., Paige, J., Zevin, B., Aggarwal, R., Grantcharov, T., … Association for Surgical Education Simulation Committee. (2015). Simulation in surgery: what's needed next? Annals of Surgery, 261(5), 846–53. http://doi.org/10.1097/SLA.0000000000000826

Steinhilber, B., Hoffmann, S., Karlovic, K., Pfeffer, S., Maier, T., Hallasheh, O., … Sievert, K. D. (2015). Development of an arm support system to improve ergonomics in laparoscopic surgery: study design and provisional results. Surgical Endoscopy and Other Interventional Techniques, 29(9), 2851–2858. http://doi.org/10.1007/s00464-014-3984-x

Thomas, W. (2006). Teaching and assessing surgical competence. Annals of the Royal College of Surgeons of England, 88(5), 429–32. http://doi.org/10.1308/003588406X116927

Tyler, J. A., Clive, K. S., White, C. E., Beekley, A. C., & Blackbourne, L. H. (2010). Current US Military Operations and Implications for Military Surgical Training. Journal of the American College of Surgeons, 211(5), 658–662. http://doi.org/10.1016/j.jamcollsurg.2010.07.009

Wiegmann, D. A., Eggman, A. A., ElBardissi, A. W., Parker, S. H., & Sundt, T. M. (2010). Improving cardiac surgical care: A work systems approach. Applied Ergonomics, 41(5), 701–712. http://doi.org/10.1016/j.apergo.2009.12.008

Williams, R. G., Klamen, D. A., & Mcgaghie, W. C. (2003). Cognitive , Social and Environmental Sources of Bias in Clinical Performance Ratings. Teaching and Learning in Medicine, 15(4), 270–292.

Williamson, R. (2004). Forgive and Remember: Managing Medical Failure, 2nd edition. Journal of the Royal Society of Medicine, 97(3), 147–148. http://doi.org/10.2307/2066542

Conclusion

This thesis advances the role of digital video review in promoting objective performance measures of surgical skill. The papers in this dissertation: (1) define a unique and consistent terminology to frame surgical skill development throughout a surgical career, (2) identify meaningful features of hand-motion associated with increasing clinician experience, (3) automatically segment a continuous video record to support on-demand review, (4) predict expert rated surgical performance in benchtop settings and (5) extrapolate performance predictions to real-world settings in the operating room. While many studies focus on measuring surgical performance using sensors or robotic interfaces, this is the first study to accomplish these aims for open surgical tasks using marker-less motion tracking of digital video. A new model of surgical skill terminology is proposed, and each paper supports traceability within the modern argument-based approaches to validity.

Increases in tenure through residency were associated with increasing movement of dominant hands, while the transition from residency to attending status was associated with reduced movement of the non-dominant hand, and less path length per cycle. This dissertation also advanced the ability to automatically deconstruct surgical video into discrete maneuvers and predict periods of suturing and tying with 79% accuracy – thereby enabling consistent predictions of cycle and completion rates for each participant. Expert rated performance was best predicted for fluidity of motion and motion economy rating scales. Both prediction models were extrapolated to video of operating room procedures and provided clear relationships between computer-predicted and expert-rated scores, albeit with an increased range of prediction for the real-world setting. The range of prediction in the operating room was similar to that of disagreement among the individual expert panelists of operating room tasks described in previous work.

The gains from each of these studies enable automatic and objective assessment of surgical performance, while providing consistent terminology and optimizing video collection, segmentation, and feedback. These abilities may culminate in future work with the design and testing of a software tool to provide formative assessment and feedback of surgical performance.

Despite these advances, the number of participants and settings involved in this study are limited. Not all experience levels completed all tasks, and the small number of retired participants, in particular, preclude findings which target the degradation of observable psychomotor skills following increased age or change in occupational role. Simulated tasks may not have been complex enough to uncover more intricate aspects of attending skill. This work also does not establish standards for objective summative assessments for competency or proficiency stages. Rather, we open the door to incorporate automatic deconstruction and performance assessment using surgical video. These results could be used for formative and ongoing quantitative assessment across all surgical roles.

Additional steps for this work include analyzing more complex simulated tasks that include shared workloads like bowel anastomoses. Promoting computer vision and video analysis of open surgical more broadly would also benefit from creation of publicly accessible, deidentified samples, similar to existing data sets available for robotic surgery. Continued recording in live operating room settings will enable additional extrapolation of performance and state predictions to real-world settings. Video capture, motion tracking, and calculation of meaningful metrics should also be integrated. These steps, while discussed here to promote study of surgical hand motion, are also widely beneficial to continuing efforts using video to detect and compare error management strategies, promote team coordination or communication, and improve operative skill through video-based coaching.

Appendices

A. Can Surgical Performance for Varying Experience be Measured from Hand Motions?

Article Citation: (Azari et al., 2018)

Azari, D. P., Greenberg, J. A., Miller, B. L., Le, B. V., Greenberg, C. C., Pugh, C. M., … Radwin, R. G. (2018). Can Surgical Performance for Varying Experience be Measured from Hand Motion? [in press]. In 2018 Annual Meeting of the Human Factors and Ergonomics Society Conference Proceedings (p. 5).

Article included starting on following page.

Can Surgical Performance for Varying Experience be Measured from Hand Motions?

David P. Azari, Brady L. Miller, Brian V. Le, Jacob A. Greenberg, Caprice C. Greenberg, Carla M. Pugh, Yu Hen Hu, and Robert G. Radwin

University of Wisconsin-Madison

This study evaluates if hand movements, tracked using digital video, can quantify in-context surgical performance. Participants of varied experience completed simple interrupted suturing and running subcuticular suturing tasks. Marker-less motion tracking software traced the two-dimensional position of a region of the hand for every video frame. Four expert observers rated 219 short video clips of participants performing the task from 0 to 10 along the following visual analog scales: fluidity of motion, motion economy, tissue handling, and coordination. Expert ratings of attending surgeon hand motions (mean=7.5, sd=1.3) were significantly greater ($p<0.05$) than medical students (mean=5.0, sd=1.9) and junior residents (mean=6.4, sd=1.5) for all rating scales. Significant differences ($p<0.02$) in mean path length per cycle were also observed both between medical students (803 mm, sd=374) and senior residents (491 mm, sd=216), and attendings (424 mm, sd=250) and junior residents (609 mm, sd=187). These results suggest that substantial gains in performance are attained after the second year of residency and that hand kinematics can predict differences in expert ratings for simulated suturing tasks commensurate with experience – a necessary step to develop valid and automatic on-demand feedback tools.

INTRODUCTION

Surgeons must develop a wide array of skills to operate effectively. The intuitive connection between skill and patient outcome is increasingly apparent (Birkmeyer et al., 2013), further reinforcing pressure to quantify and document proficiency prior to operating on patients (Aggarwal & Darzi, 2006). Assessing surgeon competency currently relies on subjective mentor observation and evaluation, in-training reports, and proxy measures such as case load and residency status (Hampton, 2015). Improving objective measures of skill is thus considered a critical step to systematically promote patient safety in the operating room (Reiley, Lin, Yuh, & Hager, 2011). The goal of this study is to examine the relationship between participant role (i.e. experience), observable hand motions, and expert-rated performance.

The most studied surgical assessment scheme – the Objective Structured Assessment of Technical Skills (OSATS) – has a strong record of valid formative feedback during training (Hatala, Cook, Brydges, & Hawkins, 2015). Implementing OSATS, however, requires real-time review and rating along a series of Likert-based scales and procedure-specific checklists. Correct application is resource intensive and time consuming (Reznick, Regehr, MacRae, Martin, & McCulloch, 1997); prompting exploration of more efficient assessment techniques such as “efficiency scores” and “snapshot assessments” (Datta, Bann, Mandalia, & Darzi, 2006). We investigate if computer vision analysis of the hands can provide a valid, automatic and more efficient measurement of surgical performance.

Motion Analysis

Hand motions (also called hand kinematics) are increasingly examined as a mode to assess live surgical performance. Our previous studies have identified differences in speed, acceleration and displacement between dominant and non-dominant hands (Glarner et al., 2014), and during live cases in the operating room by role (attendings, residents), task (tying, suturing), and varying tissue types (Frasier et al., 2016). There is also ongoing interest in representing surgical hand-motion patterns using computer automation (Ahmidi et al., 2015); and through metrics that seem to change with experience, such as “slowing down” (Moulton, Regehr, Lingard, Merritt, & MacRae, 2010), efficiency (Azari et al., 2015), entropy (Mackenzie et al., 2016), and path length (Aggarwal et al., 2007). We hypothesize that these kinds of motion attributes may be used to measure expert-rated performance along a continuous scale of experience.

Visual-Analog Scales

Previously, we created and tested subjective rating scales for expert review of short video clips (5 to 30 seconds) of open procedures (Azari et al., 2017). We utilized the existing OSATS (i.e. respect for tissue, time and motion, and instrument handling) and the Global Operative Assessment of Laparoscopic Skills (GOALS, see Vassiliou et al., 2005) as assessment blueprints to create the following visual-analog scales ranging from 0-10: (1) fluidity of motion; (2) motion economy; (3) tissue handling; (4) coordination; (5) guidance; and (6) difficulty (see Figure 1). For the current study, experts rated performance along the first four scales using a custom computer program, while participants rated their own performance along all scales immediately after the task.

METHOD

Participants and setting

Thirty-seven participants were recruited via departmental list serves and announcements during resident

training sessions. Medical students (n=10), junior residents (n=5), senior residents (n= 10), attending surgeons (n=10), and retired surgeons (n=2), agreed to have their hands recorded while completing three simple interrupted stitches, followed by a running subcuticular stitch. Junior residents were those who had completed up to three post-graduate years (PGY).

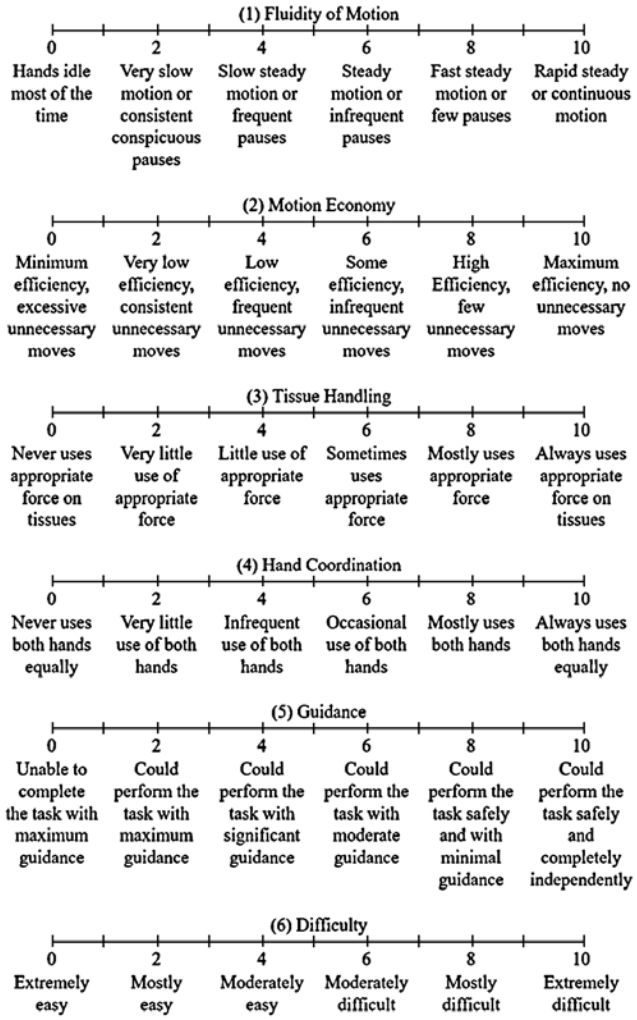


Figure 1. Series of visual-analog scales used for expert (1-4) and participant retrospective (1-6) review, adapted from the Objective Structured Assessment of Technical Skills (OSATS) and Global Operative Assessment of Laparoscopic Skills (GOALS).

Before suturing, each participant was asked to share their current training status (i.e. medical student, resident, attending, retired), years in position, and estimated case volume. Surgeon recruitment and participation was approved by the University of Wisconsin Social and Behavioral Health and Science Institutional Review Board. Written and informed consent was obtained from all participants. Interaction between participants was minimal.

Digital video cameras were positioned to preserve anonymity, observing only the participants hands and working area (Figure 2). Standard size notecards enabled calibration of the video position record into physical units (i.e. mm). Recording began after participants reviewed and signed the consent agreement. Following each task, participants self-rated their performance along each of the visual analog scales (Figure 1).

The incisions were simulated in 10.16cm x 10.16cm allevyn hydrocellular foam dressings. A scalpel was used to cut two incisions – one for each task – approximately 8cm in length and ½ the depth of the foam. Dressings were attached to 15.2cm x 15.2cm wood blocks for stability. A paper towel was folded and placed between the dressing and the wood block, stretching the foam so the interior of the incision would be visible.

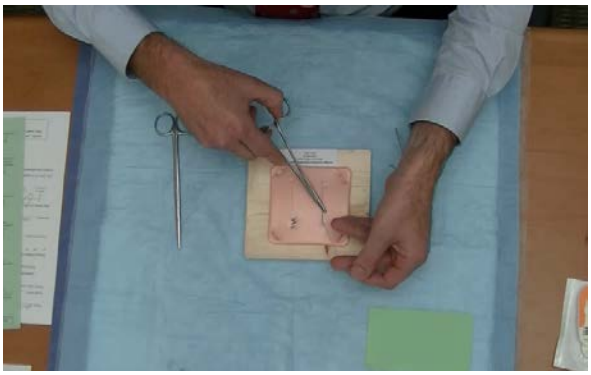


Figure 2. Overhead camera view of standard suturing station.

Motion Tracking

Clips of several active cycles (i.e. stitches) spanning 20 to 80 seconds were extracted from each video. The first suturing cycle of medical student and resident activity was treated as an acclimation period and omitted from expert rating. All available cycles of attending and retired surgeons were included for review, due to their smaller numbers. In total, 219 clips were extracted from both simple interrupted suturing (n=85) and running subcuticular sewing (n=134) tasks.

We previously developed marker-less motion tracking software to trace the position of a region of interest (ROI) across successive video frames. This software was created in Matlab and C#, utilizing the open-source OpenCVSharp (.Net wrapper for the OpenCV) vision libraries. No sensors or unique features, other than the color and shape of a participant's hand, are required to initialize the tracking program (Chen, Hu, & Radwin, 2015). Each frame of video produces a unique two-dimensional position as the ROI changes with the participant's hand. Distinct ROI's were created for each of the participant's hands, and defined to include at least two knuckles, ensuring minimal migration in the ROI when the hand changed shape. An analyst initiated the tracking algorithm and provided manual corrections as necessary.

The tracked record of the ROI location provides a rich spatiotemporal record. From the hand position at each frame, it is possible to calculate instantaneous displacement, speed, and acceleration. Additional measures compared by participant role (Table 1) include cycle frequency, path length per cycle, and jerk – the third derivative of position with respect to time.

Video Review

Four expert reviewers rated each video along the first four measures in Figure 1. The videos were randomly presented via a custom software applet (Figure 3), distributed on USB stick. Each reviewer completed a calibration activity prior to rating, in which they watched several benchmark examples and compared any discrepancies between their hypothetical rating and a previously determined panel consensus (see Azari et al., 2017).

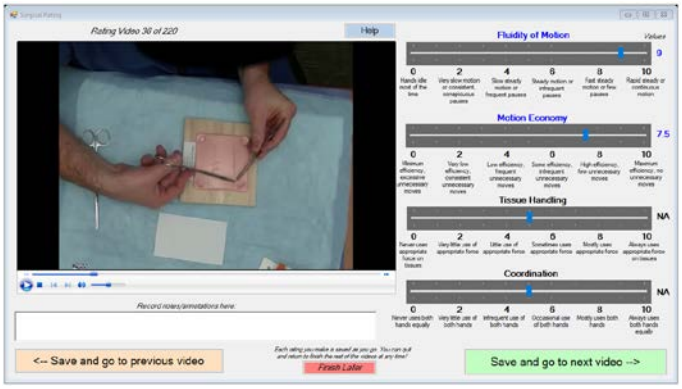


Figure 3. Standalone USB program interface for expert review and rating of surgical clips.

Rating Differences

Interrater reliability was assessed using the intraclass correlation coefficient (ICC) assuming a two-way mixed-effects model with multiple raters. Both absolute and consistency measures are reported. For clarity, single-rater coefficients are also reported. Median ratings among experts were used to generate the averages for each sample.

Differences among experience levels for each scale were tested using One-Way ANOVA and Tukey's Honestly Significant Difference (HSD) tests, given indications of normal distribution via the Shapiro-Wilks test. In the event data was non-normally distributed, Kruskal-Wallis tests were used to analyze significant differences, with pair-wise Wilcox tests utilizing the Benjamini and Hochberg p-value correction for multiple comparisons, to examine differences between groups.

RESULTS

This section describes how hand motion and expert and self-ratings of performance vary by role and skill levels.

Hand motion

Average dominant-hand kinematic results are provided in Table 1. Speed, acceleration, and jerk tended to increase as medical students became residents, and peak prior to becoming attendings. Retired surgeons exhibited slower speed and acceleration on average, but also a smaller path length per cycle. Path length per cycle of medical students differed significantly (p<0.02) from attendings, senior residents, and retired surgeons.

Expert Ratings

Intraclass correlation coefficients (ICC's) revealed good reliability among the means (absolute, consistency) of the four raters for fluidity of motion (0.83,0.85). motion economy (0.82,0.84), coordination (0.77,0.81), with poorer reliability for tissue handling (0.69,0.69). However, there was less reliability among individual panelists (absolute, consistency) for each of fluidity of motion (0.55,0.58), motion economy (0.52,0.57), tissue handling (0.35,0.36), and coordination (0.46,0.51).

Mean expert and self-ratings are summarized in Figure 4. Attending surgeon ratings were higher than all other groups for all scales, but there were no significant differences between attending surgeon and senior resident ratings. Due to the small sample size, retired surgeon ratings were omitted from significance testing.

Table 1. Observed mean kinematics.

	n	Speed (mm/s)	Accl. (mm/s^2)	Jerk (mm/s^3)	CF (hz)	PLC (mm)
Role						
MS	22	70.04	554.42	178.34	0.10	802.57
JR	10	71.71	597.54	197.51	0.14	609.15
SR	20	78.17	651.09	216.41	0.18	490.67
AT	20	71.52	587.50	191.62	0.18	423.87
RT	4	56.74	474.19	160.26	0.14	300.91
Task						
SI	36	74.32	633.04	206.24	0.20	407.04
RS	36	68.48	541.00	178.99	0.11	702.05
Role by Task						
MS-SI	9	70.11	585.14	188.85	0.15	539.23
MS-RS	9	63.53	483.91	154.65	0.07	1054.25
JR-SI	5	72.87	615.72	197.45	0.18	463.82
JR-RS	5	70.55	579.36	197.57	0.10	754.48
SR-SI	10	77.77	669.32	220.52	0.23	357.34
SR-RS	10	78.57	632.86	212.29	0.13	624.00
AT-SI	10	78.01	667.73	217.14	0.23	337.47
AT-RS	10	65.04	507.26	166.10	0.13	510.28
RT-SI	2	61.11	537.03	180.55	0.18	266.56
RT-RS	2	52.36	411.35	139.98	0.09	335.26

MS, Medical student. JR, Junior resident. SR, Senior resident. AT, Attending surgeon. RT, Retired surgeon. SI, Simple interrupted suturing. RS, Running subcuticular suturing. Accl., Acceleration. CF, Cycle frequency. PLC, Path length per cycle.

Fluidity of motion. Clips of attending surgeons were rated as more fluid (mean=7.1, sd=1.5) for all groups (p<0.05) other than senior residents (mean=6.8, sd=1.5). Medical students (mean=4.1, sd=1.9), were rated as less fluid than all other groups (p<0.04). Senior residents outperformed junior residents (mean=5.7, sd=1.6, p<0.03).

Motion economy. Medical student ratings (mean=4.3, sd=1.8) were significantly lower (p<0.01) than those for junior residents (mean=5.8, sd=1.6), senior residents (mean=6.8, sd =

1.4), and attendings (mean=7.0, sd=1.3). Senior residents and attending surgeons were rated similarly (p=0.29).

Tissue Handling. Differences (p<0.01) were observed between attendings (mean=7.8, sd=0.9) and both medical students (mean=5.9, sd=1.4) and junior residents (mean=6.9, sd=1.1). While senior residents (mean=7.7, sd=0.9) were rated higher than junior residents (p=0.01), there was no difference between attendings and senior residents (p=0.38).

Coordination. Attending surgeon coordination ratings (mean=7.8, sd=1.0) and senior resident ratings (mean=7.6, sd=1.2) were higher than those for medical students (mean=5.6, sd=1.9, p<0.02), but not for senior residents (p=0.22).

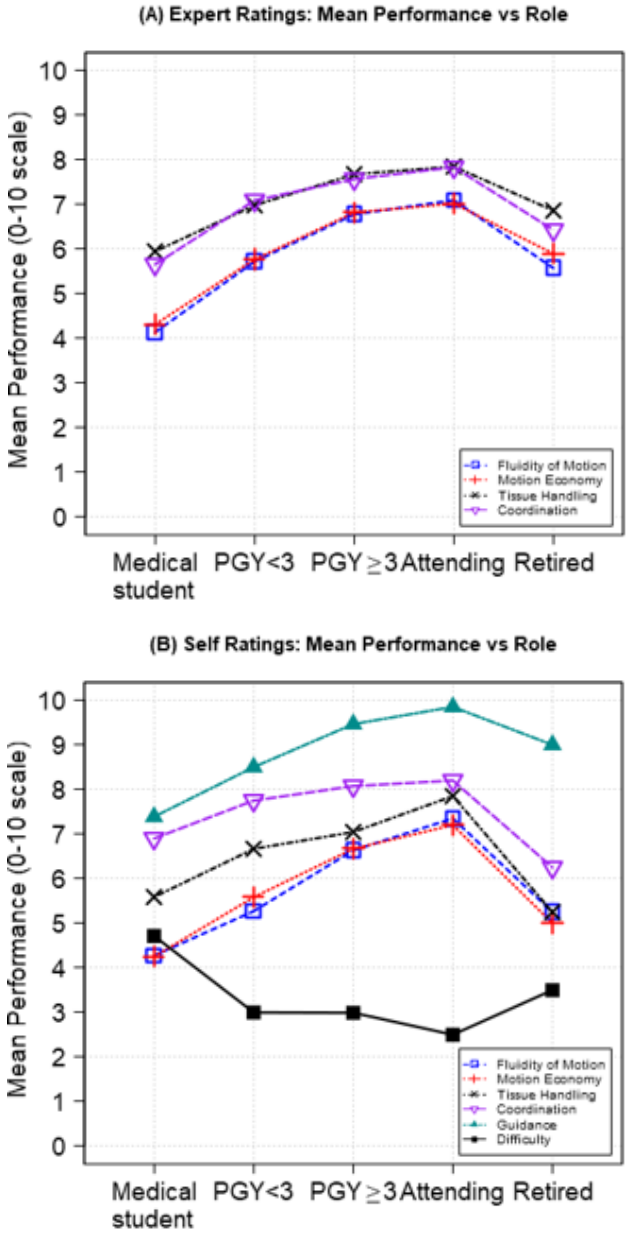


Figure 4. Mean performance on a 0 to 10 scale by participant role for expert ratings (A) and self-ratings (B).

Self-Ratings

Motion economy. Medical student self-ratings (mean=4.3, sd=22) were significantly lower (p<0.01) than senior residents (mean=6.7, sd=11) and attending surgeons (mean=7.2, sd=20), but not junior residents (mean=5.6, sd=1.1, p=0.08).

Fluidity. Attendings rated their own fluidity (mean=7.4, sd=2.0) higher than medical students (mean=4.3, sd=2.0, p<0.01) and junior residents (mean=5.3, sd=1.0, p<0.01), but not senior residents (mean=6.6, sd=1.3, p>0.10). Senior residents (mean=6.6, sd=1.3) rated themselves higher than junior residents (p<0.01).

Tissue handling. Attendings (mean=7.9, sd=1.4) and senior residents (mean=7.0, sd=1.5) rated their performance higher than medical students (mean=5.6, sd=1.8, p<0.02). There were no other significant differences by group.

Coordination. Kruskal-Wallis tests did not indicate any significant differences in self-ratings by experience group.

Guidance. Medical students (mean=7.4, sd=2.1) rated themselves lower than the combined residents (mean=9.2, sd=1.4, p=0.03), and attending surgeons (mean=9.9, sd=0.5, p=0.03).

Difficulty. There were no significant differences between difficulty ratings for residents (mean=3.0, sd=1.0) and attendings (mean=2.5, sd=1.9). Medical students, however, rated difficulty significantly higher than both residents (p<0.02) and attendings (p<0.15).

DISCUSSION

This study builds on previous work testing visual-analog rating scales of surgical performance. Recurring significant differences in expert ratings between junior and senior residents suggest that observable performance develops significantly following the second PGY. In turn, it may be possible to measure expert-rated performance as hand motion develops with experience. Self-ratings, however, among junior and senior residents only exhibited differences for the motion economy, fluidity and guidance scales. This indicates that residents are not recognizing or interpreting the same differences in coordination and tissue handling that experts are observing in their performance.

While some differences in rating are consistently pronounced (consider medical students vs senior residents and junior residents vs attendings, for example) there was little difference in expert rating between senior residents and attending surgeons. This suggests that the selected tasks may not be robust enough to measure differences as residents prepare to graduate. Skills developed after residency may only be visibly distinct during more complex cases or during procedures involving friable tissues – tasks which require more intricate features of attendings' practiced abilities like advanced cognitive decision making and efficient planning (Madani et al., 2017). Performance in these domains may be indicative of the lower path length observed for attendings, despite similar cycle frequencies between the two groups.

Still, significant differences in hand motion observed by role in this study are limited. Additional motion measures such as curvature and idle time may further add to measuring performance. Applying cyclic measurements of hand motion

(i.e. path length per cycle) to measure performance, however, depends on automatically identifying distinct stages of a task such as tying a knot, reaching for a new suture, and driving a needle – avenues of future work.

The techniques in this study do not rely on markers or sensors but require a clear, consistent, and non-obstructed view of the hands. Finding a camera position with minimal distortion and minimal occlusion from head movement is challenging. Even with ideal camera placement, the surgeon's head occasionally occludes the hands, and all surgeons had periods where their hands leave the frame. These instances are processed manually, slowing down our ability to translate hand motion into a useful data record. Extensive time was also allocated to reviewing and ensuring that the tracking result matched the hand location for all frames, and that motions from any extraneous behaviors were excluded. Medical students would occasionally pause to ask a question, or otherwise make a statement in which they used their hands to gesture. These periods were manually identified and removed.

There are also several audio-visual challenges to scaling up this kind of study, including correct frame identification, frame-rate (or dropped frame) and video codec conversion, compression, and calibration. These challenges can be overcome, given certain filming and software settings, but they pose a significant hurdle to wide-spread adoption and consistent review in healthcare. The required materials, on the other hand, are readily accessible (i.e. webcams, foam dressings, video software), and could be widely distributed.

Future work will focus on expanding the available motion metrics, exploring the relationship between self and expert ratings by task type, and using significant kinematic measures identified in this study to automatically predict expert ratings across the range of experience and observed performance.

CONCLUSIONS

This study utilized digital video and computer vision of hand motion during simulated suturing tasks to examine the relationship between hand kinematics and performance ratings. Experts rated senior residents and attending surgeons consistently higher than medical students and junior residents for motion economy, fluidity of motion, tissue handling, and coordination. Statistically significant differences in ratings and hand motions were discovered for varying experience groups. Fluidity of motion and path length per cycle were the most distinct measures of participant performance. These results suggest that computer vision of hand motion can predict differences in expert ratings for simulated suturing tasks commensurate with experience, enabling valid and automatic on-demand feedback tools for surgical training and coaching.

ACKNOWLEDGEMENTS

The authors acknowledge the following persons for their assistance to this work: Trenton Feda, Akshat Khanna, Calvin Kwan, Reginald Bruskewitz, Barb Lewis.

REFERENCES

Aggarwal, R., & Darzi, A. (2006). Technical-skills training in the 21st century. The New England Journal of Medicine, 355(25), 2695–2696. https://doi.org/10.1056/NEJMe068179

Aggarwal, R., Grantcharov, T., Moorthy, K., Milland, T., Papasavas, P., Dosis, A., … Darzi, A. (2007). An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Annals of Surgery, 245(6), 992–999.

Ahmidi, N., Poddar, P., Jones, J. D., Vedula, S. S., Ishii, L., Hager, G. D., & Ishii, M. (2015). Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. International Journal of Computer Assisted Radiology and Surgery, 10(6), 981–991. https://doi.org/10.1007/s11548-015-1194-1

Azari, D. P., Frasier, L. L., Quamme, S. R. P., Greenberg, C. C., Pugh, C. M., Greenberg, J. A., & Radwin, R. G. (2017). Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. Annals of Surgery, XX(Xx), 1. https://doi.org/10.1097/SLA.0000000000002478

Azari, D. P., Pugh, C. M., Laufer, S., Kwan, C., Chen (Eric), C.-H., Yen, T. Y., … Radwin, R. G. (2015). Evaluation of Simulated Clinical Breast Exam Motion Patterns Using Marker-Less Video Tracking. Human Factors, 58(3), 427–440. https://doi.org/10.1177/0018720815613919

Birkmeyer, J. D., Finks, J. F., O'Reilly, A., Oerline, M., Carlin, A. M., Nunn, A. R., … Birkmeyer, N. J. O. (2013). Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine, 369(15), 1434–42. https://doi.org/10.1056/NEJMsa1300625

Chen, C.-H., Hu, Y. H., & Radwin, R. G. Video Hand-Tracking In The Industrial Working Environment. 40th IEEE International Conference on Acoustics, Speech and Signal Processing. Brisbane, Australia, April 2015.

Datta, V., Bann, S., Mandalia, M., & Darzi, A. (2006). The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. American Journal of Surgery, 192(3), 372–378.

Frasier, L. L., Azari, D. P., Ma, Y., Quamme, S. R. P., Radwin, R. G., Pugh, C. M., … Greenberg, C. C. (2016). A marker-less technique for measuring kinematics in the operating room. Surgery. https://doi.org/10.1016/j.surg.2016.05.004

Glarner, C. E., Hu, Y. Y., Chen, C. H., Radwin, R. G., Zhao, Q., Craven, M. W., … Greenberg, C. C. (2014). Quantifying technical skills during open operations using video-based motion analysis. Surgery (United States), 156(3), 729–734. https://doi.org/10.1016/j.surg.2014.04.054

Hampton, T. (2015). Efforts Seek to Develop Systematic Ways to Objectively Assess Surgeons' Skills. Jama, 313(8), 782–784. https://doi.org/10.1001/jama.2015.233

Hatala, R., Cook, D. A., Brydges, R., & Hawkins, R. (2015). Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. Advances in Health Sciences Education, 20(5), 1149–1175. https://doi.org/10.1007/s10459-015-9593-1

Mackenzie, C. F., Watts, D., Patel, R., Yang, S., Hagegeorge, G., Hu, P. F., … Tisherman, S. (2016). Sensor-free Computer-Vision hand-motion entropy and video-analysis of technical performance during open surgery on fresh cadavers: report of methodology and analysis. In Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting (pp. 691–695).

Madani, A., Vassiliou, M. C., Watanabe, Y., Al-Halabi, B., Al-Rowais, M. S., Deckelbaum, D. L., … Feldman, L. S. (2017). What Are the Principles That Guide Behaviors in the Operating Room? Annals of Surgery, 265(2), 255–267. https://doi.org/10.1097/SLA.0000000000001962

Moulton, C. A., Regehr, G., Lingard, L., Merritt, C., & MacRae, H. (2010). Slowing down to stay out of trouble in the operating room: remaining attentive in automaticity. Acad Med, 85(10), 1571–1577.

Reiley, C. E., Lin, H. C., Yuh, D. D., & Hager, G. D. (2011). Review of methods for objective surgical skill evaluation. Surgical Endoscopy and Other Interventional Techniques, 25(2), 356–366.

Reznick, R., Regehr, G., MacRae, H., Martin, J., & McCulloch, W. (1997). Testing technical skill via an innovative "bench station" examination. American Journal of Surgery, 173(3), 226–230.

Vassiliou, M. C., Feldman, L. S., Andrew, C. G., Bergman, S., Leffondré, K., Stanbridge, D., & Fried, G. M. (2005). A global assessment tool for evaluation of intraoperative laparoscopic skills. American Journal of Surgery, 190(1),

B. Software Rating Program

This program is written in C#, and accomplishes the following:

1. Loads window to prompt for user name (demographic).
2. Loads (randomly) all videos saved in video folder.
3. Saves all user changes in summary file saved to “sessions” folder.

Upon opening, the program saves a “session” with the username and date. This is also the name of a csv file, stored in a “sessions” folder which includes video order, and number of completed ratings after each “save and…” selection made by the user. All user interaction are saved in a “backups” folder for error tracing. The user can return to complete un-rated videos at a later time, given that they choose to open the previously saved “session” from the first screen.

Program.cs

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Threading.Tasks;
using System.Windows.Forms;
namespace v3
{
static class Program
{
/// <summary>
/// The main entry point for the application.
/// </summary>
[STAThread]
static void Main()
{
Application.EnableVisualStyles();
Application.SetCompatibleTextRenderingDefault(false);
Application.Run(new Demographic());
}
}
}
```

Demographic.cs

```
using System;
using System.IO;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data;
using System.Drawing;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Windows.Forms;

namespace v3
{
   public partial class Demographic : Form
   {
      public static string userName;
      public static string csvSessionToRead;
      public static bool continuePreviousExperiment;
      public static string mainPath =
Path.GetFullPath(System.IO.Directory.GetCurrentDirectory());//MAIN PATH


      public Demographic()
      {
         Console.WriteLine("Main path is:" + mainPath);

         InitializeComponent();
         prevExperiment.Enabled = false; //Set initial state of prev exp button to false, so can't
load up anything
         Console.WriteLine("demographic activated");


         //Populate list of CSVs
         string[] csvSessionFiles = Directory.GetFiles(mainPath + "\\sessions\\",
"*.csv*").Select(Path.GetFileName).ToArray();
         for(int f = 0; f < csvSessionFiles.Length; f++)
         {
            listBox1.Items.Add(csvSessionFiles[f]);
            Console.WriteLine(csvSessionFiles[f] + " added to box.");
         }
      }

      private void StartExperiment_Click(object sender, EventArgs e)
      {
         //string filePath = Environment.GetFolderPath(Environment.SpecialFolder.Desktop);
```

```
    //string fullSaveFolderPath = filePath + "\\TestCSVWrite\\" + nameField.Text + ".csv";
    userName = nameField.Text;
    Console.WriteLine("userName" + userName);

    continuePreviousExperiment = false; //Start with clean slate...


    RatingApplet R = new RatingApplet();
    R.Show();
    this.Hide();
    Console.WriteLine("Start Experiment Button Clicked");
}

private void prevExperiment_Click(object sender, EventArgs e)
{
    continuePreviousExperiment = true; //Continue where we left off...
    userName = nameField.Text;
    Console.WriteLine("userName" + userName);

    RatingApplet R = new RatingApplet();
    R.Show();
    this.Hide();
    Console.WriteLine("Cont Prev Experiment Button Clicked");
}



private void Demographic_Load(object sender, EventArgs e)
{
    nameField.Text = Environment.UserName;

    Console.WriteLine("demographic load");
}

private void name_TextChanged(object sender, EventArgs e)
{
    userName = nameField.Text;
    Console.WriteLine(userName);
}

private void userNamePrompt_Click(object sender, EventArgs e)
{
    userName = nameField.Text;
    Console.WriteLine(userName);
}


private void listBox1_SelectedIndexChanged_1(object sender, EventArgs e)
```

```
        {
            if (listBox1.SelectedItem == null)
            {
                prevExperiment.Enabled = false; //Disallow button click

                //DO NOTHING
                Console.WriteLine("NOTHING DONE");
                listBox1.ClearSelected();

            } else //SAVE CSV FILE FOR LATER READ IN
            {
                prevExperiment.Enabled = true; //allow button click

                csvSessionToRead = mainPath + "\\sessions\\" + listBox1.SelectedItem.ToString();
                Console.WriteLine(csvSessionToRead);
            }
        }

        private void label2_Click(object sender, EventArgs e)
        {

        }
    }
}
```

Demographic.Designer.cs

```
using System;
using System.IO;

namespace v3
{
   partial class Demographic
   {
      /// <summary>
      /// Required designer variable.
      /// </summary>
      private System.ComponentModel.IContainer components = null;

      /// <summary>
      /// Clean up any resources being used.
      /// </summary>
      /// <param name="disposing">true if managed resources should be disposed; otherwise,
false.</param>
      protected override void Dispose(bool disposing)
      {
         if (disposing && (components != null))
         {
            components.Dispose();
         }
         base.Dispose(disposing);
      }

      #region Windows Form Designer generated code

      /// <summary>
      /// Required method for Designer support - do not modify
      /// the contents of this method with the code editor.
      /// </summary>
      private void InitializeComponent()
      {
         this.nameField = new System.Windows.Forms.TextBox();
         this.StartExperiment = new System.Windows.Forms.Button();
         this.userNamePrompt = new System.Windows.Forms.Label();
         this.prevExperiment = new System.Windows.Forms.Button();
         this.listBox1 = new System.Windows.Forms.ListBox();
         this.label1 = new System.Windows.Forms.Label();
         this.label2 = new System.Windows.Forms.Label();
         this.SuspendLayout();
         //
         // nameField
         //
```

```
        this.nameField.Font = new System.Drawing.Font("Microsoft Sans Serif", 12F,
System.Drawing.FontStyle.Regular, System.Drawing.GraphicsUnit.Point, ((byte)(0)));
        this.nameField.Location = new System.Drawing.Point(252, 20);
        this.nameField.Margin = new System.Windows.Forms.Padding(2);
        this.nameField.Name = "nameField";
        this.nameField.Size = new System.Drawing.Size(167, 26);
        this.nameField.TabIndex = 0;
        this.nameField.Text = "<username>";
        this.nameField.TextAlign = System.Windows.Forms.HorizontalAlignment.Center;
        this.nameField.TextChanged += new System.EventHandler(this.name_TextChanged);
        //
        // StartExperiment
        //
        this.StartExperiment.Font = new System.Drawing.Font("Microsoft Sans Serif", 14.25F,
System.Drawing.FontStyle.Regular, System.Drawing.GraphicsUnit.Point, ((byte)(0)));
        this.StartExperiment.Location = new System.Drawing.Point(41, 67);
        this.StartExperiment.Margin = new System.Windows.Forms.Padding(2);
        this.StartExperiment.Name = "StartExperiment";
        this.StartExperiment.Size = new System.Drawing.Size(378, 75);
        this.StartExperiment.TabIndex = 1;
        this.StartExperiment.Text = "Begin new rating session";
        this.StartExperiment.UseVisualStyleBackColor = true;
        this.StartExperiment.Click += new System.EventHandler(this.StartExperiment_Click);
        //
        // userNamePrompt
        //
        this.userNamePrompt.AutoSize = true;
        this.userNamePrompt.Font = new System.Drawing.Font("Microsoft Sans Serif", 12F,
System.Drawing.FontStyle.Regular, System.Drawing.GraphicsUnit.Point, ((byte)(0)));
        this.userNamePrompt.Location = new System.Drawing.Point(37, 23);
        this.userNamePrompt.Name = "userNamePrompt";
        this.userNamePrompt.Size = new System.Drawing.Size(192, 20);
        this.userNamePrompt.TabIndex = 2;
        this.userNamePrompt.Text = "Please confirm username:";
        this.userNamePrompt.Click += new System.EventHandler(this.userNamePrompt_Click);
        //
        // prevExperiment
        //
        this.prevExperiment.Font = new System.Drawing.Font("Microsoft Sans Serif", 14.25F,
System.Drawing.FontStyle.Regular, System.Drawing.GraphicsUnit.Point, ((byte)(0)));
        this.prevExperiment.Location = new System.Drawing.Point(41, 387);
        this.prevExperiment.Name = "prevExperiment";
        this.prevExperiment.Size = new System.Drawing.Size(378, 65);
        this.prevExperiment.TabIndex = 3;
        this.prevExperiment.Text = "Continue previous session";
        this.prevExperiment.UseVisualStyleBackColor = true;
        this.prevExperiment.Click += new System.EventHandler(this.prevExperiment_Click);
        //
```

```
        // listBox1
        //
        this.listBox1.FormattingEnabled = true;
        this.listBox1.Location = new System.Drawing.Point(41, 248);
        this.listBox1.Name = "listBox1";
        this.listBox1.Size = new System.Drawing.Size(378, 121);
        this.listBox1.TabIndex = 4;
        this.listBox1.SelectedIndexChanged += new
System.EventHandler(this.listBox1_SelectedIndexChanged_1);
        //
        // label1
        //
        this.label1.AutoSize = true;
        this.label1.Font = new System.Drawing.Font("Microsoft Sans Serif", 12F,
System.Drawing.FontStyle.Regular, System.Drawing.GraphicsUnit.Point, ((byte)(0)));
        this.label1.Location = new System.Drawing.Point(27, 212);
        this.label1.Name = "label1";
        this.label1.Size = new System.Drawing.Size(413, 20);
        this.label1.TabIndex = 5;
        this.label1.Text = "Select a session below, and click on \"continue\" to resume";
        //
        // label2
        //
        this.label2.AutoSize = true;
        this.label2.Font = new System.Drawing.Font("Microsoft Sans Serif", 12F,
System.Drawing.FontStyle.Regular, System.Drawing.GraphicsUnit.Point, ((byte)(0)));
        this.label2.Location = new System.Drawing.Point(192, 180);
        this.label2.Name = "label2";
        this.label2.Size = new System.Drawing.Size(59, 20);
        this.label2.TabIndex = 6;
        this.label2.Text = "~ OR ~";
        this.label2.Click += new System.EventHandler(this.label2_Click);
        //
        // Demographic
        //
        this.AutoScaleDimensions = new System.Drawing.SizeF(6F, 13F);
        this.AutoScaleMode = System.Windows.Forms.AutoScaleMode.Font;
        this.ClientSize = new System.Drawing.Size(461, 473);
        this.Controls.Add(this.label2);
        this.Controls.Add(this.label1);
        this.Controls.Add(this.listBox1);
        this.Controls.Add(this.prevExperiment);
        this.Controls.Add(this.userNamePrompt);
        this.Controls.Add(this.StartExperiment);
        this.Controls.Add(this.nameField);
        this.Margin = new System.Windows.Forms.Padding(2);
        this.Name = "Demographic";
        this.Text = "Demographic";
```

```
            this.Load += new System.EventHandler(this.Demographic_Load);
            this.ResumeLayout(false);
            this.PerformLayout();

        }

        #endregion

        private System.Windows.Forms.TextBox nameField;

        private System.Windows.Forms.Button StartExperiment;
        private System.Windows.Forms.Label userNamePrompt;
        private System.Windows.Forms.Button prevExperiment;
        private System.Windows.Forms.ListBox listBox1;
        private System.Windows.Forms.Label label1;
        private System.Windows.Forms.Label label2;
    }
}
```

Main Form

```
using System;
using System.IO;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data;
using System.Drawing;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Windows.Forms;
using System.Text.RegularExpressions;
using System.Net.Mail;//1116

namespace v3
{
   public partial class RatingApplet : Form
   {
      //PRELOAD DEMOGRAPHIC FORM:
      //The first screen to collect information and setup experient
      Demographic userData = new Demographic();

      //DECLARE THINGS
      public int pageIndex = 0; //Will follow pages (i.e. videos)
      public int btnClick = 0; //Will follow clicks of "save results" (i.e. rows in the tracking
spreadsheet...

      bool emailAvailed = false;

      string oldFileName;
      string copyFileName;

      string sessionName = (Demographic.userName + "_" +
DateTime.Now.ToString("yyyyMMddhhmmssfff") + ".csv");

      string driveLetter = Demographic.mainPath.Split(':')[0]; // NEW FOR 20180327
      string trueUserName = Demographic.userName; // REPLACED W. PREVIOUS IS
LOADED

      string sessionPathFolder;
      string sessionPathFull; //Holds the path for ALL RESULTS (timeseries of interactions)

      string[] videoFiles; //Contains all the paths for all videos within a defined directory...

      //string vFile; //Container for my own stupid programming mistakes :(

      Random rnd = new Random(); //Enables random function below
```

```
    int[] randOrder; //Contains the randomized order of all videos to be loaded...

    int numComp = 0; //Initialized just in case open and don't finish anything

    //RATING ARRAYS
    string[] videoFilePaths;

    int[] rateS1;
    int[] rateS2;
    int[] rateS3;
    int[] rateS4;

    string[] annot;//Store annotations for each video
    string[] vidCompleteYrN;

    List<string> lines = new List<string>();

    public RatingApplet()
    {
       InitializeComponent();
       TopMost = true;
       selectVideoSet();//MOVED FROM AXMEDIA PLAYER
       loadVideo(pageIndex);//MOVED FROM AXMEDIA PLAYER
       resetSliderValues(pageIndex);
       updateProgressBar(pageIndex);

    }

    private int[] getRandomOrder(int numberOfVideoFiles)
    {
       int[] randomOrder = new int[numberOfVideoFiles];//Initialize random order array of size
#videos!

       for (int k = 0; k < numberOfVideoFiles; k++)
       {
          randomOrder[k] = -1;
       }
       int i = 0;
       while (i < numberOfVideoFiles)
       {
          var ii = rnd.Next(0, numberOfVideoFiles);
          if (!randomOrder.Contains(ii))
          {
             randomOrder[i] = ii;
             i++;
          }
       }
       return randomOrder;
```

```
    }

    private void loadVideo(int videoNumberToLoad)
    {
       Console.WriteLine("Drive letter is: " + driveLetter);

       string fileHolderTemp;
       //Console.WriteLine("annotation herez: " + annot[videoNumberToLoad]);
       annotations.Text = annot[videoNumberToLoad]; //Update annotation

       if (Demographic.continuePreviousExperiment == true)
       {
          fileHolderTemp = videoFilePaths[videoNumberToLoad].Split(':')[1];
          videoPlayer.URL = string.Concat(driveLetter + ":" + fileHolderTemp);
           //videoPlayer.URL = videoFilePaths[videoNumberToLoad]; //REMOVED FOR
TESTING
           //vFile = videoFilePaths[videoNumberToLoad];
           //Console.WriteLine("LoadVid: " + videoFilePaths[videoNumberToLoad]);
           Console.WriteLine("LoadVid: " + string.Concat(driveLetter, ":", fileHolderTemp));
       } else {

           fileHolderTemp = videoFilePaths[randOrder[videoNumberToLoad]].Split(':')[1];
           videoPlayer.URL = string.Concat(driveLetter + ":" + fileHolderTemp);
           //videoPlayer.URL = videoFilePaths[randOrder[videoNumberToLoad]];
           //vFile = videoFilePaths[randOrder[videoNumberToLoad]];
           Console.WriteLine("LoadVid: " + string.Concat(driveLetter, ":", fileHolderTemp));

       }

          // Console.WriteLine("RandVid: " + randOrder[videoNumberToLoad]);
           Console.WriteLine("BtnClck: " + btnClick);
           Console.WriteLine("PageInd: " + videoNumberToLoad);
           Console.WriteLine("SliderV: " + (((double)slider1.Value) / 10).ToString());
           Console.WriteLine("RaterS1: " + rateS1[videoNumberToLoad]);

    }

    private void selectVideoSet()
    {
       DialogResult closing = MessageBox.Show(
           "Please watch each video and drag the sliders accordingly. Any changes you make to
the slider positions are saved as you go." + Environment.NewLine + Environment.NewLine +
           "SAVE AND CONTINUE LATER: Since your progress is saved, feel free to skip over
individual videos or navigate with the 'go to next' and 'go to back' buttons, or quit at any time." +
Environment.NewLine + Environment.NewLine +
           "You can return later, and as many times as you wish in order to finish previous
sessions (listed on the previous page). You won't have to re-rate any videos if you load a
previous session." + Environment.NewLine + Environment.NewLine +
```

```
            "If you have trouble, please let David know: [Contact Information Redacted]." +
            Environment.NewLine +
            sessionPathFull,
            //+ Environment.NewLine ,
            //+ Environment.NewLine + "Would you like to close the program now?"),
            "Thank you for taking the time to participate!",
           MessageBoxButtons.OK);

        if (Demographic.continuePreviousExperiment == true)
        {

            sessionName = Demographic.csvSessionToRead;
            trueUserName = sessionName.Split('\\').Last();
            trueUserName = trueUserName.Split('_').First(); // FOR DISPLAY ONLY

            MessageBox.Show("Program will resume previous session, " + trueUserName + "
from the first incomplete video. ");

            string[] words = sessionName.Split('\\');
            oldFileName = words[words.Length-1];
            Console.WriteLine("select video set: oldFileName = " + oldFileName);
            string[] temp = oldFileName.Split('-');
            sessionName = temp[temp.Length - 1];
            oldFileName = Demographic.mainPath + "\\sessions\\" + oldFileName;

            copyFileName = Demographic.mainPath + "\\backups\\" + "RELOADED - " +
sessionName + DateTime.Now.ToString("yyyyMMddhhmmssfff") + ".csv";
            System.IO.File.Copy(oldFileName, copyFileName);

            sessionPathFolder = Demographic.mainPath + "\\output\\";
            sessionPathFull = sessionPathFolder + sessionName;

            //Console.WriteLine("TRUE SESSION NAME ADAPTITION!" + sessionName);

            //Console.WriteLine("REACHED NEW SECTION, BOOYAH");

            var path = @Demographic.mainPath;

            List<string> retro_vidPath = new List<string>();
            List<int> retro_vidOrdr = new List<int>();
            List<int> retro_c01 = new List<int>();
            List<int> retro_c02 = new List<int>();
            List<int> retro_c03 = new List<int>();
            List<int> retro_c04 = new List<int>();
            List<string> retro_annot = new List<string>();
            List<string> retro_vidComplete = new List<string>();
```

```
        //string[] compSwitch;

        using (var reader = new StreamReader(Demographic.csvSessionToRead))
        {
           //int f = 0;
           while (!reader.EndOfStream)
           {
              var line = reader.ReadLine();
              var values = line.Split(',');

              //compSwitch = values[0].Split('\\'); //ADDED S.T. ONLY pATH AFTER
INITIAL LETTER
              retro_vidPath.Add(values[0]);
              //retro_vidPath.Add(compSwitch[2]);
              //retro_vidOrdr.Add(Convert.ToInt32(values[1]));
              retro_c01.Add(Convert.ToInt32(values[1]));
              retro_c02.Add(Convert.ToInt32(values[2]));
              retro_c03.Add(Convert.ToInt32(values[3]));
              retro_c04.Add(Convert.ToInt32(values[4]));
              retro_vidComplete.Add(values[5]);
              retro_annot.Add(values[6]);

              Console.WriteLine("should see the annotation here: " + values[6]);

           }
        };
        //UPDATE TRACKER ARRAYS TO MATCH LOADED CSV FILE
        videoFilePaths = retro_vidPath.ToArray();
        randOrder = Enumerable.Range(0,videoFilePaths.Length).ToArray();
        rateS1 = retro_c01.ToArray();
        rateS2 = retro_c02.ToArray();
        rateS3 = retro_c03.ToArray();
        rateS4 = retro_c04.ToArray();
        vidCompleteYrN = retro_vidComplete.ToArray();
        annot = retro_annot.ToArray();
        videoFiles = videoFilePaths;

        //After storing all variables into arrays, delete!
        //System.IO.File.Delete(oldFileName);

        Console.Write(vidCompleteYrN);

        //NEW 1114 update to last completed :D
        int v = 0;
        while (vidCompleteYrN[v] == "y")
        {
```

```
            Console.WriteLine(vidCompleteYrN[v] + " is entry in vidComplete arry ind of " +
v);
              v++;

              if(v + 1 > vidCompleteYrN.Length)
              {
                 //v = v - 1;
                Console.WriteLine("BOOYAAAA TIS OK - SHOULD LOAD #7!");
                pageIndex = vidCompleteYrN.Length-1; Console.WriteLine("v pg ind = " +
pageIndex);
                 emailDavid();
                 break;
              } else
              {
                 pageIndex = v;
              }
            }

        }
        else
        {
           //DEFINE VIDEO DIRECTORY
           string videoFilesPath = (Demographic.mainPath + "\\videos\\");
           Console.WriteLine(("Video Directory: " + videoFilesPath));
           videoFilePaths = Directory.GetFiles(videoFilesPath);


           //RANDOMIZE LOAD ORDER OF FOR ALL DIRECTORY VIDEO FILES
           randOrder = getRandomOrder(videoFilePaths.Length);
           Console.WriteLine(randOrder);

           //INITIALIZE ALL THE SLIDER STORAGE ARRAYS
           rateS1 = Enumerable.Repeat(-1, videoFilePaths.Length).ToArray();
           rateS2 = Enumerable.Repeat(-1, videoFilePaths.Length).ToArray();
           rateS3 = Enumerable.Repeat(-1, videoFilePaths.Length).ToArray();
           rateS4 = Enumerable.Repeat(-1, videoFilePaths.Length).ToArray();
           annot = Enumerable.Repeat(String.Empty, videoFilePaths.Length).ToArray();
           vidCompleteYrN = Enumerable.Repeat("n", videoFilePaths.Length).ToArray();
        }

        //MOVED UP FROM BELOW! Hope this helps...
        sessionPathFolder = (Demographic.mainPath + "\\output\\");
        sessionPathFull = (Demographic.mainPath + "\\output\\" + sessionName); //Session
names will not overwrite, as they are uniquely dated.


    }
```

```
        private void emailDavid()
        {
            string mailto = string.Format("mailto:{0}?Subject={1}&attachment={2}&Body={3}",
                "[contact info redacted]",
                "COMPLETED "+sessionName,
                sessionPathFull,
                Demographic.userName + " has completed all ratings. Please arrange to pick up the
USB.    " + DateTime.Now.ToString("yyyyMMddhhmmssfff"));

            if (!emailAvailed) //IF HAVEN"T OPENED WINDOW YET, DO SO. Otherwise, do
nothing!
            {
              System.Diagnostics.Process.Start(mailto);
            }

            emailAvailed = true;
        }

        private void axWindowsMediaPlayer1_Enter(object sender, EventArgs e)
        {
            int temp = 1; //Just check for ALL COMPLETED --> replaced temp = pageIndex
            Console.WriteLine("reached axMedia main loop - page index is " + pageIndex + " temp
is" + temp);
            //TRYING TO MOVE THINGS AROUND 1134, moved from AXMediaPlayer
MENACE - tried at 11:44
            while (vidCompleteYrN[temp] == "y")
            {
                temp++; Console.WriteLine("reached AXMedia while loop temp val: " + temp);
                if (temp + 1 > vidCompleteYrN.Length)
                {
                    MessageBox.Show(("All videos in this session have been rated! You can still review
these videos, if you wish. Otherwise, the USB is ready to be returned! Please email David at
[contact info redacted]"));
                    pageIndex = vidCompleteYrN.Length - 1;
                    Console.WriteLine("About to break...");
                    break;
                }
            }
            Console.WriteLine("ENDED axMedia main loop - page index is " + pageIndex + " temp
is" + temp);


            //First Time Session Output Formation

            if (!Directory.Exists(sessionPathFolder))
            { //CHECK to ensure that the directory exists - if not, create it!
                Directory.CreateDirectory(sessionPathFolder);
                Console.WriteLine("Output Directory Created!");
```

```
            MessageBox.Show(("New output directory (to save all future ratings) created in: " +
sessionPathFolder), "Output directory created.");
        }


        //Output PREEMPTIVE!
        string[][] output = new string[][]{
              new string[]{
                 ("Timestamp"),//DateTime.Now.ToString("yyyyMMddhhmmssfff"),
                 ("Username"),//Demographic.userName,
                 ("ButtonClx"),
                 ("PageIndex"),//("PgInd: " + videoNumberToLoad),
                 //("RandIndex"),//Num: " + randOrder[videoNumberToLoad].ToString()),
                 ("VideoPath"),//videoFiles[randOrder[videoNumberToLoad]],
                 (slider1.Name),//(((double)slider1.Value)/10).ToString(),//COULD REPLACE
THESE DIRECT CALLS WITH ARRAY REFERENCES....
                 (slider2.Name),//(((double)slider2.Value)/10).ToString(),
                 (slider3.Name),//(((double)slider3.Value)/10).ToString(),
                 (slider4.Name),//(((double)slider4.Value)/10).ToString(),
                 //(slider5.Name),//(((double)slider5.Value)/10).ToString(),
                 //(slider6.Name),//(((double)slider5.Value)/10).ToString(),
                 ("Annotations")//annot[videoNumberToLoad]
              }//End new string
           };

        if (pageIndex < 1) //Page Index starts at 0
        {
           //WRITE OUT BIG FILE
           int length = output.GetLength(0);
           StringBuilder sb = new StringBuilder();
           for (int i = 0; i < length; i++)
           {
              sb.AppendLine(string.Join(",", output[i]));
           }
           File.AppendAllText(sessionPathFull, sb.ToString());//SAVE FILE
        }
        //BY DEFAULT

     }

     private void saveValues(int videoNumberToLoad)
     {
        //On Every Save Define new file name based on completion rate
        string moveFileName = Demographic.mainPath + "\\backups\\" + "from - " +
sessionName + "  " + numComp.ToString() + " of " +
        videoFilePaths.Length.ToString() + "  " +
DateTime.Now.ToString("yyyyMMddhhmmssfff") + ".csv";
        Console.WriteLine(moveFileName + " created as movefilename");
```

```
        if (File.Exists(oldFileName))
        {
           System.IO.File.Copy(oldFileName, moveFileName);
        }
        //annot[videoNumberToLoad] =
annotations.Text.Replace(System.Environment.NewLine, " ");
        annot[videoNumberToLoad] = Regex.Replace(annotations.Text, @"\r\n?|\n|,", " ");

        string[][] output = new string[][]{
              new string[]{
                 DateTime.Now.ToString("yyyyMMddhhmmssfff"),
                 Demographic.userName,
                 (""+btnClick),
                 (""+videoNumberToLoad),
                 //(randOrder[videoNumberToLoad].ToString()),
                 videoFilePaths[randOrder[videoNumberToLoad]],

(""+rateS1[videoNumberToLoad]),//(((double)slider1.Value)/10).ToString(),//COULD
REPLACE THESE DIRECT CALLS WITH ARRAY REFERENCES....
                 (""+rateS2[videoNumberToLoad]),//(((double)slider2.Value)/10).ToString(),
                 (""+rateS3[videoNumberToLoad]),//(((double)slider3.Value)/10).ToString(),
                 (""+rateS4[videoNumberToLoad]),//(((double)slider4.Value)/10).ToString(),
                 //(((double)slider5.Value)/10).ToString(),
                 //(((double)slider6.Value)/10).ToString(),
                 annot[videoNumberToLoad]
              }//End new string
                };
        int length = output.GetLength(0);
        StringBuilder sb = new StringBuilder();
        for (int i = 0; i < length; i++)
        {
           sb.AppendLine(string.Join(",", output[i]));
        }
        Console.WriteLine("FULL PATH IS: " + sessionPathFull);
        File.AppendAllText(sessionPathFull, sb.ToString());//SAVE FILE



        //On Every Save, Update Completion Rate - resave CSV
        numComp = 0;
        for (int v = 0; v < videoFilePaths.Length; v++)
        {
           if (rateS1[v] < 0) { vidCompleteYrN[v] = "n"; }
           else if (rateS2[v] < 0) { vidCompleteYrN[v] = "n"; }
           else if (rateS3[v] < 0) { vidCompleteYrN[v] = "n"; }
           else if (rateS4[v] < 0) { vidCompleteYrN[v] = "n"; }
           else { vidCompleteYrN[v] = "y"; numComp++; }
```

```
            }
        double compRate = numComp / videoFilePaths.Length;

        //OVERWRITE NEW FILE NAME AFTER UPDATE
        string newFileName = Demographic.mainPath + "\\sessions\\" + numComp.ToString() +
" of " +
            videoFilePaths.Length.ToString() + " complete -" + sessionName;

        //Create NEW file it if exists...
        if (File.Exists(newFileName))
        {
            File.WriteAllText(newFileName, String.Empty);
        }
        //string newFileName = Demographic.mainPath + "\\sessions\\" + sessionName;

        for (int v = 0; v < videoFilePaths.Length; v++)
        {
            using (StreamWriter sw = new StreamWriter(newFileName, true))
            {
                sw.WriteLine(string.Join(",", videoFilePaths[randOrder[v]].ToString(),
                                        //randOrder[v].ToString(),
                                        rateS1[v].ToString(),
                                        rateS2[v].ToString(),
                                        rateS3[v].ToString(),
                                        rateS4[v].ToString(),
                                        vidCompleteYrN[v].ToString(),
                                        annot[v].ToString().Replace(",","-")));

            }
        }

        if (File.Exists(oldFileName)) //If old file exists
        {
            if(oldFileName != newFileName) //And that old file is different than the new one
which was just written
            {
                System.IO.File.Delete(oldFileName);
            }
        }
        oldFileName = newFileName; //Set old for next go around...
    }

    private void updateProgressBar(int videoNumberToLoad)
    {
        string vidCompletionRate = "";
        if(vidCompleteYrN[videoNumberToLoad] == "y")
        {
            vidCompletionRate = (trueUserName
```

```
            + " Rating Video " + (videoNumberToLoad + 1) + " of " + videoFilePaths.Length
            + " complete.");

        } else
        {
            vidCompletionRate = (trueUserName
            + " Rating Video " + (videoNumberToLoad + 1) + " of " + videoFilePaths.Length);
        }
        StatusLabel.Text = vidCompletionRate;
        annotations.Text = annot[videoNumberToLoad];
    }

    private void resetSliderValues(int videoNumberToLoad)
    {
        annotations.Text = annot[videoNumberToLoad];

        //CHECK IF NOT SAVED, Fill w. Defaults, Otherwise, load active data
        //S1
        if (rateS1[videoNumberToLoad] == -1)
        {
            slider1.Value = 50; label1.Text = "NA";
            label1.ForeColor = System.Drawing.Color.Black; label7.ForeColor =
System.Drawing.Color.Black;
        } else
        {
            slider1.Value = rateS1[videoNumberToLoad];
            label1.Text = (((double)slider1.Value) / 10).ToString();
            label1.ForeColor = System.Drawing.Color.Blue; label7.ForeColor =
System.Drawing.Color.Blue;
        }
        //S2
        if (rateS2[videoNumberToLoad] == -1)
        {
            slider2.Value = 50; label2.Text = "NA";
            label2.ForeColor = System.Drawing.Color.Black; label8.ForeColor =
System.Drawing.Color.Black;
        }
        else
        {
            slider2.Value = rateS2[videoNumberToLoad];
            label2.Text = (((double)slider2.Value) / 10).ToString();
            label2.ForeColor = System.Drawing.Color.Blue; label8.ForeColor =
System.Drawing.Color.Blue;
        }
        //S3
        if (rateS3[videoNumberToLoad] == -1)
        {
            slider3.Value = 50; label3.Text = "NA";
```

```
            label3.ForeColor = System.Drawing.Color.Black; label9.ForeColor =
System.Drawing.Color.Black;
         }
         else
         {
            slider3.Value = rateS3[videoNumberToLoad];
            label3.Text = (((double)slider3.Value) / 10).ToString();
            label3.ForeColor = System.Drawing.Color.Blue; label9.ForeColor =
System.Drawing.Color.Blue;
         }
         //S4
         if (rateS4[videoNumberToLoad] == -1)
         {
            slider4.Value = 50; label4.Text = "NA";
            label4.ForeColor = System.Drawing.Color.Black; label10.ForeColor =
System.Drawing.Color.Black;
         }
         else
         {
            slider4.Value = rateS4[videoNumberToLoad];
            label4.Text = (((double)slider4.Value) / 10).ToString();
            label4.ForeColor = System.Drawing.Color.Blue; label10.ForeColor =
System.Drawing.Color.Blue;
         }
         updateProgressBar(videoNumberToLoad); //Call progress bar from here...
      }

      private void prevVideo_Click(object sender, EventArgs e)
      {
         saveValues(pageIndex);
         if (pageIndex > 0)
         {

            pageIndex = pageIndex - 1; //If not at first page, decrement, load associated video, and
reset slider values
            //MessageBox.Show("Saved values. PI Decremented one to " + pageIndex);
            loadVideo(pageIndex); //Load associated video
            resetSliderValues(pageIndex);
         }
         else
         {
            pageIndex = 0; //RESET TO BASE VALUE
            MessageBox.Show("This is the first video in the set.", "No previous video to load.");
         }
         resetSliderValues(pageIndex);
         btnClick = btnClick + 1; //btnClick is for ALL EVENTS
      }//End Previous Video
```

```
    private void button1_Click(object sender, EventArgs e)//GO TO NEXT VIDEO
    {
        saveValues(pageIndex);

        pageIndex = pageIndex + 1;

        //CHECK FOR ENDING CONDITION, AND, IF SO, CAN CLOSE THE FORM
        if ((pageIndex + 1) > videoFilePaths.Length)
        {
            //MessageBox.Show(("Thank you for your participation! You can now close the rating
program. Results are saved here: " + sessionPathFull), "Experiment Complete!");
            pageIndex = videoFilePaths.Length - 1;
            //resetSliderValues(pageIndex);
            Console.WriteLine("pageIndex reset to max videos: " + pageIndex);
            //POPUP MESSAGE
            DialogResult closing = MessageBox.Show(("Thank you for your participation! Feel
free to go back and review your ratings. You can also load this session later to finish any videos
you skipped."
                                                    //+ Environment.NewLine + Environment.NewLine +"Results
are saved here: " + sessionPathFull
                                                    + Environment.NewLine + Environment.NewLine +
                                                    "You have completed " + numComp.ToString() + " of " +
videoFilePaths.Length + " videos."
                                                    + Environment.NewLine + Environment.NewLine +
                                                    "Would you like to close the program now?"), "Reached last
video!",
                                                   MessageBoxButtons.YesNo);

            string lastFileName = Demographic.mainPath + "\\backups\\" + "REACHED THE
END " + " -from- " + sessionName + " " + numComp.ToString() + " of " +
            videoFilePaths.Length.ToString() + " curTime is " +
DateTime.Now.ToString("yyyyMMddhhmmssfff") + ".csv";
            Console.WriteLine(lastFileName + " created as lastFileName");
            if (File.Exists(oldFileName))
            {
                System.IO.File.Copy(oldFileName, lastFileName);
            }

            //SEE IF NEED TO EMAIL!
            if (vidCompleteYrN.Contains("n"))
            {
                //DO NOTHING
            } else {
                emailDavid();

            }
```

```
            if (closing == DialogResult.Yes)//CLOSE IF DESIRED!
            {
               Application.Exit(); //Exit v3
               this.Close(); //CLOSE PROGRAM CORRECTLY
               System.Diagnostics.Process.Start(@sessionPathFolder); //Open up folder to view
saved CSV
            }
         }
         else//Otherwise
         {
            resetSliderValues(pageIndex);
            loadVideo(pageIndex);
         } //End Latter IF statement chain
         btnClick = btnClick + 1;
      }//End button click function :)

      //SLIDER REAL TIME VALUES SEE
      private void slider1_Scroll(object sender, EventArgs e)
      {
         rateS1[pageIndex] = (slider1.Value); //Save Value!
         label1.Text = (((double)slider1.Value) / 10).ToString();
         label1.ForeColor = System.Drawing.Color.Blue;
         label7.ForeColor = System.Drawing.Color.Blue;
      }
      private void slider2_Scroll(object sender, EventArgs e)
      {
         rateS2[pageIndex] = (slider2.Value); //Save Value!
         label2.Text = (((double)slider2.Value) / 10).ToString();
         label2.ForeColor = System.Drawing.Color.Blue;
         label8.ForeColor = System.Drawing.Color.Blue;
      }
      private void slider3_Scroll_1(object sender, EventArgs e)
      {
         rateS3[pageIndex] = (slider3.Value); //Save Value!
         label3.Text = (((double)slider3.Value) / 10).ToString();
         label3.ForeColor = System.Drawing.Color.Blue;
         label9.ForeColor = System.Drawing.Color.Blue;
      }
      private void slider4_Scroll_1(object sender, EventArgs e)
      {
         rateS4[pageIndex] = (slider4.Value); //Save Value!
         label4.Text = (((double)slider4.Value) / 10).ToString();
         label4.ForeColor = System.Drawing.Color.Blue;
         label10.ForeColor = System.Drawing.Color.Blue;
      }

      private void RatingApplet_Load(object sender, EventArgs e)
      {
```

```
    TopMost = false;
  }
  private void label13_Click(object sender, EventArgs e)
  {

  }
  private void label15_Click(object sender, EventArgs e)
  {

  }
  private void label37_Click(object sender, EventArgs e)
  {

  }
  private void label1_Click(object sender, EventArgs e)
  {

  }

  private void StatusLabel_Click(object sender, EventArgs e)
  {

  }

  private void RatingApplet_FormClosing(object sender, FormClosingEventArgs e)
  {
    saveValues(pageIndex);

    //MessageBox.Show("Thank you! All ratings are saved in the session: " +
    //    sessionName);

    Application.Exit();

   }

  private void button1_Click_1(object sender, EventArgs e)
  {
    DialogResult closing = MessageBox.Show(
            "Please watch each video and drag the sliders accordingly. Any changes you
make to the slider positions are saved as you go." + Environment.NewLine +
Environment.NewLine +
            "SAVE AND CONTINUE LATER: Since your progress is saved, feel free to
skip over individual videos or navigate with the 'go to next' and 'go to back' buttons, or quit at
any time." + Environment.NewLine + Environment.NewLine +
            "You can return later, and as many times as you wish in order to finish previous
sessions (listed on the previous page). You won't have to re-rate any videos if you load a
previous session." + Environment.NewLine + Environment.NewLine +
            "If you have trouble, please let David know: [contact info redacted] " +
```

```
                    sessionPathFull,
                    //+ Environment.NewLine ,
                    //+ Environment.NewLine + "Would you like to close the program now?"),
                    " Thank you for taking the time to participate!",
                   MessageBoxButtons.OK);
        }

        private void label33_Click(object sender, EventArgs e)
        {

        }

        private void label30_Click(object sender, EventArgs e)
        {

        }

        private void label58_Click(object sender, EventArgs e)
        {

        }

        private void label58_Click_1(object sender, EventArgs e)
        {

        }

        private void button2_Click(object sender, EventArgs e)
        {
           saveValues(pageIndex);

           MessageBox.Show("Thank you! All ratings are saved in the session: " +
              sessionName); //DONT NEED THIS - ALREADY DOES ON CLOSING

           Application.Exit();
        }

        private void label23_Click(object sender, EventArgs e)
        {

        }
    } // Partial Class
} //End Namespace
```