AWARD NUMBER: W81XWH-15-2-0030

TITLE: Robotic Surgery Readiness (RSR): A Prospective Randomized Skills Decay Recognition and Prevention Study

PRINCIPAL INVESTIGATOR: Thomas Lendvay

CONTRACTING ORGANIZATION: University of Washington

REPORT DATE: OCTOBER 2020

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release; Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
OCTOBER 2020	FINAL	15Jul2015 – 14July 2020

4. TITLE AND SUBTITLE

Robotic Surgery Readiness (RSR): A Prospective Randomized Skills Decay Recognition and Prevention Study

5a. CONTRACT NUMBER
W81XWH-15-2-0030

5b. GRANT NUMBER
12195004

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Dr. Thomas Lendvay
Email: Thomaslendvey@seattlechildrens.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Washington
4333 Brooklyn Ave NE
Seattle, WA 98105-1016

8. PERFORMING ORGANIZATION REPORT

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release, Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Importance: Intra-operative errors correlate with surgeon skill and skill declines with intervals of inactivity

Objective: The goals of this research were to identify the optimal virtual reality (VR) warm-up curriculum to prime a surgeon's technical skill and validate benefit in the operating room.

Design: Surgeons were randomized to receive six trial sessions of a designated set of VR modules on the da Vinci Skills Simulator. After performing their curricula, warm-up effect was assessed based on performance on a criterion task. The optimal warm-up curriculum was chosen from the group with the best task time and video review-based technical skill. Robot-assisted surgery (RAS)-experienced surgeons were then recruited to either receive or not receive warm-up before surgery. Skill in the first 15 minutes of surgery was assessed by blinded surgeon and crowd worker review as well as tool motion metrics.

Setting: Surgeons and trainees were recruited from urology, gynecology and general surgery from military, Veteran's Administration, public academic, and private medical centers.

Participants: Forty-one surgeons participated in the warm-up curriculum selection, while 34 practicing RAS-experienced surgeons participated in intra-operative warm-up validation.

Intervention: The intervention was performing VR warm-up prior to human RAS.

Main Outcome and Measure: Warm-up effect was measured using objective performance metrics and video review using the Global Evaluative Assessment of Robotic Skills (GEARS) tool. Linear mixed effects models with a random intercept for each surgeon and nonparametric modified Friedman tests were used for analysis.

Results: A Running Suture task on the simulator was on average 31.3 seconds faster, and had the highest GEARS scores, thus was chosen as the optimal curriculum. Thereafter, surgeons completed 343 surgeries with corresponding video and tool motion data. No statistically significant differences in skill were observed with the warm-up intervention.

Conclusion and Relevance: We conclude that a robotic VR warm-up prior to performing the early stages of surgery does not impact the technical skill of the surgeon.

15. SUBJECT TERMS
Robotic Surgery, Readiness, da Vinci Simulator, Virtual reality, simulation curriculum, GEARS – Global Evaluation Assessment of Robotic Skills, Surgical Education

16. SECURITY CLASSIFICATION OF:			**17. LIMITATIO**	**18. NUM**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT U	**b. ABSTRAC T** U	**c. THIS PAGE** U	UU	16	USAMRDC **19b. TELEPHONE NUMBER** *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

1. **Introduction…….…….…….3**
2. **Keywords……..…….…….4**
3. **Accomplishments……..…….…….…….4**
4. **Impact…… 18**
5. **Changes/Problems...…….…….…….…… 19**
6. **Products…….……..…..…….21**
7. **Participants & Other Collaborating Organizations…..……... 23**
8. **Special Reporting Requirements…….…… 23**
9. **Appendices…….……... 24**

INTRODUCTION

Medical errors account for 98,000 deaths each year, atop a higher number of patient complications. An estimated one third of these errors are surgical and the average American can expect to have 7 surgeries in their lifetime. Although surgical simulation's role in off-line teaching is accepted, simulation may also have a role in identifying decrements in surgical performance and offer a system for mitigating skill decay. Perez et al. concluded that skill may decay at different degrees depending on the type of skill, and that training to mitigate decay may require targeted curricula based on the specific decay signatures. These methods were described for laparoscopic surgery skills, yet no warm-up has been studied for robot-assisted surgery (RAS) which has seen considerable adoption in the United States with the creation of the da Vinci surgical robot by Intuitive Surgical (Sunnyvale, CA).

While several other checklist-type preparation procedures exist in the operating room (OR), there is still no metric which measures how prepared a surgeon is following periods of inactivity. In other professions, after participating in a warm-up, task performance can be elevated. Previous work additionally shows that a warm-up decrement (WUD) exists, in which the effects of warm-up begin decaying approximately 30 minutes after performing a warm-up task. These hypotheses have already found four factors which most affect WUD: (1) intervals between practice, (2) complexity and type of task, (3) strategies employed to maintain skills, and (4) individual differences. Virtual reality (VR) proficiency-based warm-up modules have been shown to combat skill decay most effectively. This research group identified in a prior study that VR warm-up primes experienced robotic surgeons when looking at task time and errors, thus we sought to assess the warm-up benefit among practicing surgeons.

Measuring surgical technical skill in an objective manner is difficult, with the gold standard being video review by expert surgeons. Previous work has shown that video review of technical skill by non-expert crowds is accurate. When compared to expert surgeon review, crowds of non-experts were able to predict the pass/fail rate of surgeons with 100% accuracy in retrospective studies. Tool motion data captured from the da Vinci has also been associated with surgical skill. Metrics such as Task Time TT), spectral arc length (SAL), camera use per minute (CUP), normalized angular displacement (NAD), rate of orientation change (ROC), and mean velocity (MV), have all been found to associate with surgical skill.

In this study we sought to find the optimal VR warm-up curriculum to properly prime surgeons for RAS and validate the warm-up benefit in the OR by measuring objective performance metrics and obtaining video review of participants.

KEYWORDS

Robotic Surgery
Readiness
da Vinci Simulator
Virtual Reality
Simulation Curriculum
GEARS - Global Evaluative Assessment of Robotic Skills
Surgical Education

ACCOMPLISHMENTS

What were the major goals of the project?

YEAR 1 (0-12 Months)

1) Study design and skill decay model construction, supplies purchasing and acquisition.
2) Set-up and flow within and between simulation centers.
3) Development of robust methods for collecting, merging, and verifying simulator, operative video, and kinematic data. Tool motion metric capturing technology development
4) Subject recruitment
5) Skills decay testing.
6) Independent video review of VR simulator criterion performances using GEARS tool.
7) Analysis of performance metrics.

Deliverables: Quantifiable performance signatures of robotic surgery skills decay assessment. Initial analysis of data. Preliminary RSR warm-up curriculum.

YEAR 2 (12-24 Months)

1) Finalize and validate RSR curriculum and benchmarks.
2) Intra-operative RSR warm-up subject recruitment.
3) RSR curriculum hypothesis testing, intra-operative data collection.
4) Independent video review of surgical performances using GEARS.

Deliverable: Finalized RSR warm-up curriculum, initial dataset and data quality assessment.

YEAR 3 (24-36 Months)

1) Continued intra-operative RSR curriculum hypothesis testing.
2) Continued independent video review of operative performances using GEARS.
3) Bio statistical analysis and model cross-validation.
4) Abstract and manuscript drafting.

Final Deliverables: Validated RSR warm-up curriculum and assessment tools. Methodology for quantifying robotic surgery skills decay. Peer-reviewed publication, presentation at national meeting.

YEAR 4 (36-48) No Cost Extension – continuation of year 3

1) DoD approved extension from July ’18 to Jan ’20

Final Deliverables: Recruited surgeons for warm-up hypothesis testing

YEAR 5 (48-54) No Cost Extension – continuation of year 3

1) Independent video review of operative performances using GEARS.
2) Bio statistical analysis and model cross-validation.
3) Abstract and manuscript drafting.

Final Deliverables: Completed, validated RSR warm-up curriculum and assessment tools. Methodology for quantifying robotic surgery skills decay. Peer-reviewed publication, presentation at national meeting.

What was accomplished under these goals?

YEAR 1 (July 2015 – July 2016)

- Study Design and Materials
    - Study design and skill decay model constructed by month three
    - Acquiring major equipment took the first year to complete
    - Computers purchased and set up for all sites (UW, MAMC, VA, FL Hospital)
    - Robust methods developed for collecting, merging, and verifying simulator operative video and kinematic data
    - Subjects given unique identifiers based on their location (identifiers were kept at each site and only de-identified data was collected by UMN team).
    - Subjects to take intake demographics questionnaire and complete proficiency training
- Set up and Flow within and between simulation centers
    - UW biostatistician helped improve data intake and form development
    - All site coordinators were trained on REDCap
    - REDCap was used to collect data and randomize subjects
    - All de-identified video and simulator data output collected by the team at UMN for analysis.
- Development of robust methods for collecting, merging, and verifying simulator, video, and optical tracking data
    - UMN reworked software to provide high definition video acquisition and compression compared to previous versions with poorer resolution and video artifacts
    - UMN successfully installed and monitored data acquisition from all sites and synchronized with central database
    - Collaborative agreement between UMN, UW, and Intuitive Surgical Inc. was on-going to extract tool motion metrics from da Vinci robot.
- Subject Recruitment
    - Aim 1 Recruitment (RSR Curriculum design) began at ¾ sites.
    - MAMC contacted 15 potential participants and consented 9 subjects. 6 completed their proficiency stage at the end of year one.
        - Acted as test site to confirm planning worked (allowed for editing of forms and procedures to make flow optimally).
    - UW/VA contacted 45 with none proficient.
    - Florida Hospital had none at this point (due to needing to reconcile contract language with UW before proceeding).
- Skills Decay Testing
    - MAMC had six ready to be randomized and begin Aim 1 by the end of Year 1.

YEAR 2 (July 2016 – July 2017)

- Set up and flow within and between simulation centers
    - Recruited subjects completed intake demographics questionnaire and begin proficiency training.
    - Subject details were kept at each site and only de-identified data was collected by the team at Minnesota (UMN). All subjects were given a unique identifier based on their location.
    - REDCap was used at all sites to collect data and to randomize the subjects.
    - UMN provided early drafts and input on revisions for the data collection forms and a UW biostatistician continued to improve the REDCap data collection process.
    - UMN reworked the software to provide high definition video acquisition and compression and successfully installed and monitored the acquisition of data from all sites. The data was

synchronized with our central database as designed. UNM continued to rectify the Google Drive videos, REDCap data, and dVLogger data to ensure all logs are complete and no data is lost and where necessary, investigated incomplete logs and proposes fixes.
    - A collaborative triparty agreement was put into place, brokered between the Office of Sponsored Research at the University of Washington, UMN and Intuitive Surgical, Inc. to provide kinematic data directly from the da Vinci robots.
- Subject Recruitment
    - Aim 1 Recruitment conducted at all four sites with consented and proficient subjects
    - Florida Hospital and UW successfully reconciled contract language which had previously delayed their active participation
- Skills Decay Testing
    - See results in Table 1 below
    - 28 subjects completed their study sessions.
    - MAMC consented 35 subjects at the end of year 2 (At this point, 20 progressed to completion, 5 continued to progress through their sessions, 8 were working to achieve proficiency, 1 withdrew, and 1 was lost to follow up).
    - UW/VA consented 28 subjects by the end of year 2 (At this point, 10 were randomized and 8 of the 10 completed all study sessions, two continued to progress through their sessions, 9 were working to achieve proficiency, 3 withdrew due to scheduling issues, and 6 were lost to follow up).
    - Florida hospital consented 4 subjects by the end of year 2 (At this point, 1 completed all study sessions, 2 withdrew due to time constraints, and 1 was relocated to another hospital).
        - Table 1: Recruitment data summary by site, n(%)

Robotic Surgery Readiness (RSR) : Enrollment Report

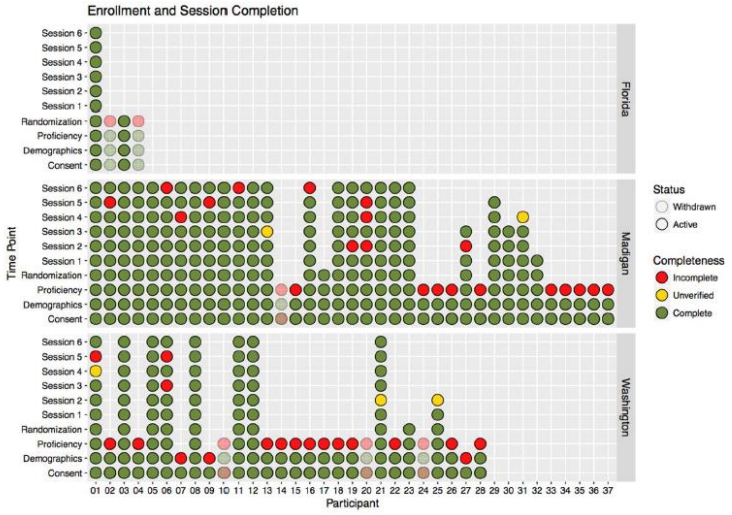
July 31, 2017

Table 1: Recruitment data summary by site, n (%)

	Florida	Madigan	Washington	Total
Target Enrollment	20	20	30	70
Total Randomized	**2 (10)**	**26 (130)**	**10 (33)**	**38 (54)**
Completed Participation	**1 (5)**	**20 (100)**	**8 (27)**	**29 (41)**
Consented*	4 (20)	37 (185)	28 (93)	69 (99)
Gender				
Female	1 (25)	9 (24)	8 (29)	18 (26)
Male	3 (75)	28 (76)	17 (61)	48 (70)
Other	0 (0)	0 (0)	0 (0)	0 (0)
Seniority				
Attending	4 (100)	11 (30)	2 (7)	17 (25)
Senior	0 (0)	9 (24)	9 (32)	18 (26)
Junior	0 (0)	17 (46)	14 (50)	31 (45)
Specialty				
Urology	0 (0)	12 (32)	11 (39)	23 (33)
Gynocology	1 (25)	8 (22)	2 (7)	11 (16)
General	2 (50)	8 (22)	10 (36)	20 (29)
Thoracic	0 (0)	0 (0)	0 (0)	0 (0)
Cardiothoracic	0 (0)	0 (0)	0 (0)	0 (0)
ENT	1 (25)	9 (24)	2 (7)	12 (17)
Handedness				
Left	1 (25)	1 (3)	1 (4)	3 (4)
Right	2 (50)	34 (92)	22 (79)	58 (84)
Ambidextrous	1 (25)	2 (5)	2 (7)	5 (7)
Musical Experience	2 (50)	22 (59)	18 (64)	42 (61)
Deployed	0 (0)	6 (16)	0 (0)	6 (9)

*Denominator for percentages below

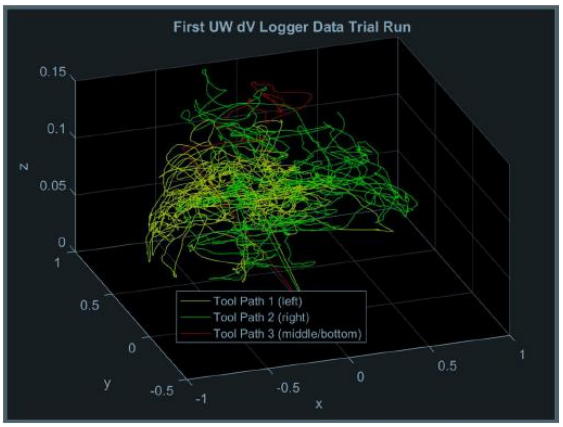
        - Figure 1: Enrollment and session completion



- Data entry was ongoing and REDCap forms were updated as needed for optimal data capture.
- Syncing issues were identified between the Florida and UMN and hardware at Florida could not be debugged remotely. The laptop was returned, debugged and mailed back to the Florida site.
- A data logging time stamp issue was identified from the dVSS. Software was created to parse data from the logger into human readable format for rectification and logging purposes. Code was modified to meet end user needs and software was subsequently created to rectify DV logger data to REDCap logs.
- We developed data collection tools for Aim 2. These tools, to be implemented on paper or through REDCap, include a participant demographics questionnaire, session data caption, and outcome data collection.
- Designed and built the REDCap database for Aim 2. This included developing a randomization schema and system to randomize each participant to the RSR assignment

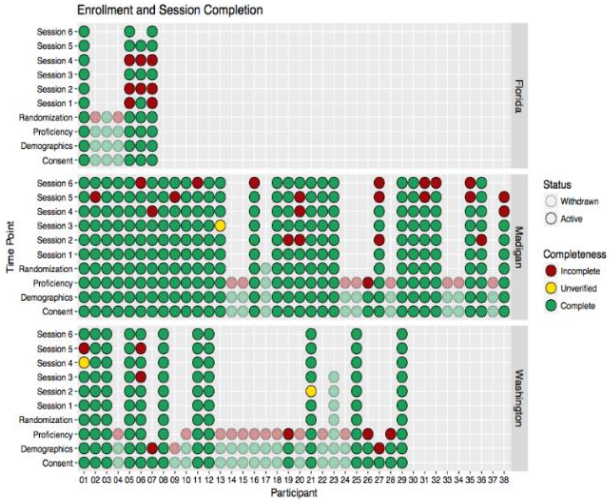
- Site Visits
  - We conducted site visits to FL Hospital and to MAMC last October. These visits were very informative. FL Hospital has a wonderful facility but its distance from the hospital and subjects was a problem. The FL Hospital team was very capable and continued to make efforts with the local surgery population to recruit them into the study. The MAMC coordinator exceeded recruitment expectations for her site.
  - In April, team members met for 3 days in Seattle for orientation to and installation of the Intuitive DV Logger. Dr. Lendvay and Lois Meryman from UW, Dr. Kowalewski and Anna French from UM, and Evelyn George from Madigan and were in attendance. Omid Mohareri from Intuitive Surgical demonstrated the data capture process. We tested integration of dVLogger with our hardware (only Apple wireless worked without flaw, hence move to Apple devices for Aim 2 data collection/control)
  - Test data was captured, reviewed by the team and verified that full data streams are available from dVLogger for extracting sample metrics like path length and working volume (see sample plot below, Figure 2)

    - Figure 2: First UW dV logger data trial run



YEAR 3 (July 2017 – July 2018)

- Aim 1 enrollment was complete. The RSR Aim 2 warm-up curriculum and benchmarks have been finalized and validated. See Figure 3 below for Aim 1 completion stats.
    - Florida Hospital enrolled 7; 4 subjects completed sessions, 2 withdrew due to time constraints, 1 was relocated to another hospital
    - MAMC enrolled 37 subjects; 28 completed sessions, 1 withdrew, and 8 were lost to follow-up
    - UW/VA enrolled 29 subjects; 11 completed sessions, 5 dropped out due to scheduling issues, and 12 were lost to follow up
    - Figure 3: Enrollment and Session Completion



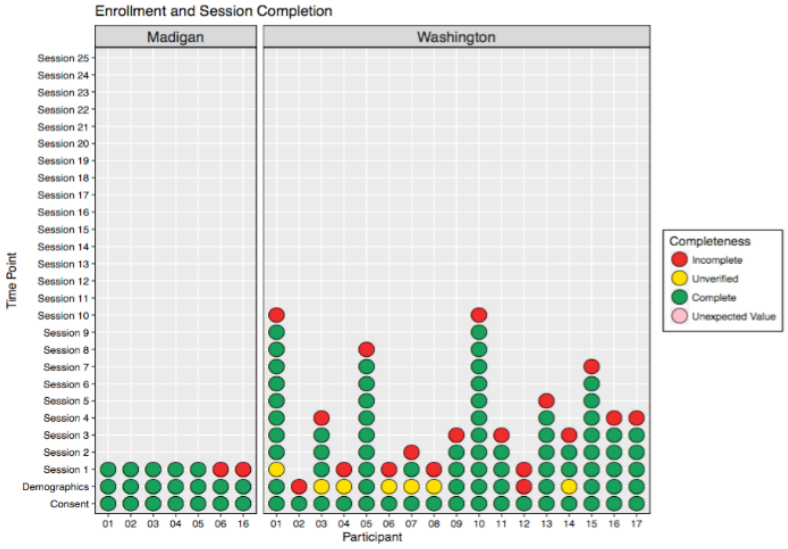
- Aim 2 Intra-operative RSR subject recruitment and hypothesis testing began March 2018 at Madigan Army Medical Center and the University of Washington/Veterans Administration Sites; Florida Hospital determined it was unable to participate in Aim 2 and thus dropped out and closed their protocol.
- Independent video review of operative performances using GEARS continues.
- Set up and flow between simulation centers
    - Work Flow
        - Recruited subjects sign consent and complete the intake demographics questionnaire. Subject details are kept at each site and only de-identified data is collected by the team at Minnesota (UMN). All subjects are given a unique identifier based on their location.
        - The data is synchronizing with our central database as designed. UNM continued to optimize REDCap data, conducts dVLogger data integrity checks to ensure all logs are complete and no data is lost and where necessary, investigates incomplete logs and proposes fixes.

- Data entry was ongoing and REDCap forms were updated as needed for optimal data capture.
- One dVLogger at the UW site was returned to Intuitive because of faulty software that led to corruption or loss of 10 cases of data. This was reconciled with a replacement dVLogger that captured data appropriately which was double checked after each session.
- Protocol was modified as needed for the implementation of Aim 2 (see BEFORE CASE CHECKLIST below).
    - BEFORE CASE CHECKLIST
        - Ethernet cable connected
        - L and R SDI cables connected
        - Power cables connected
        - Confirm stereo configurations on tower
        - Confirm USB connection between hard drive and dV-Logger
        - Power on dV-Logger well before surgeon begins operating
        - Log into web-UI
        - Access webpage via http://10.42.0.1
        - Check hard drive storage space
        - Confirm that data is streaming
        - Confirm that recording has started before the procedure begins
    - If any items from the checklist were not addressed, please describe why:

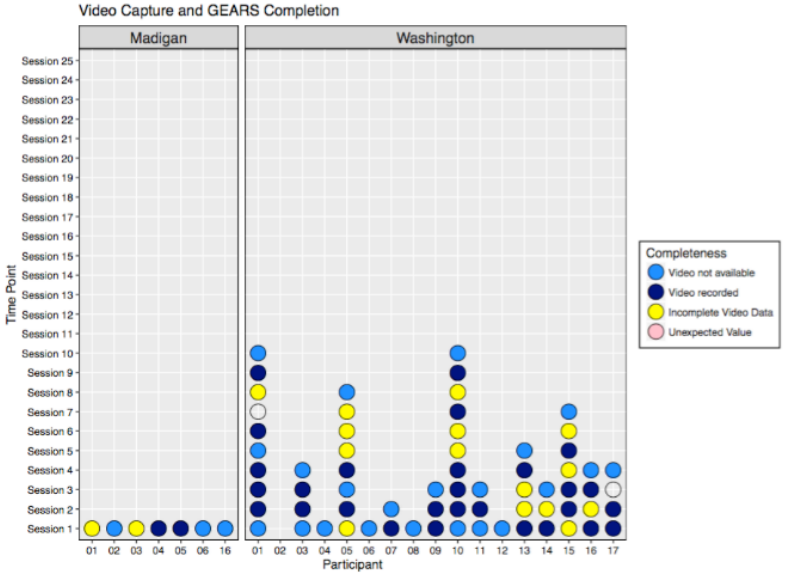
- Aim 2 Recruitment (see Table 2)
    - UW/VA consented 17 surgeons. One withdrew. The remaining surgeons continue to progress through sessions.
    - MAMC consented 7 surgeons. Two are no longer based at this site. The remaining surgeons continue to progress through sessions.
    - Table 2 Recruitment data summary by site, n(%)

	Madigan	Washington	Total
Target Enrollment	6	20	26
Randomized Sessions	**7**	**67**	**74**
Completed Sessions	**5**	**51**	**56**
Consented*	7 (117)	17 (85)	24 (92)
Gender			
Female	0 (0)	5 (29)	5 (21)
Male	7 (100)	10 (59)	17 (71)
Other	0 (0)	0 (0)	0 (0)
Specialty			
Urology	2 (29)	8 (47)	10 (42)
Gynocology	0 (0)	5 (29)	5 (21)
General	4 (57)	1 (6)	5 (21)
Thoracic	0 (0)	0 (0)	0 (0)
Cardiothoracic	0 (0)	0 (0)	0 (0)
Handedness			
Left	0 (0)	0 (0)	0 (0)
Right	6 (86)	15 (88)	21 (87)
Ambidextrous	0 (0)	0 (0)	0 (0)
Musical Experience	2 (29)	4 (24)	6 (25)
Deployed	4 (57)	0 (0)	4 (17)

- Figure 4: Aim 2 Enrollment and Session completion



- Figure 5: Aim 2 Video capture and gears completion



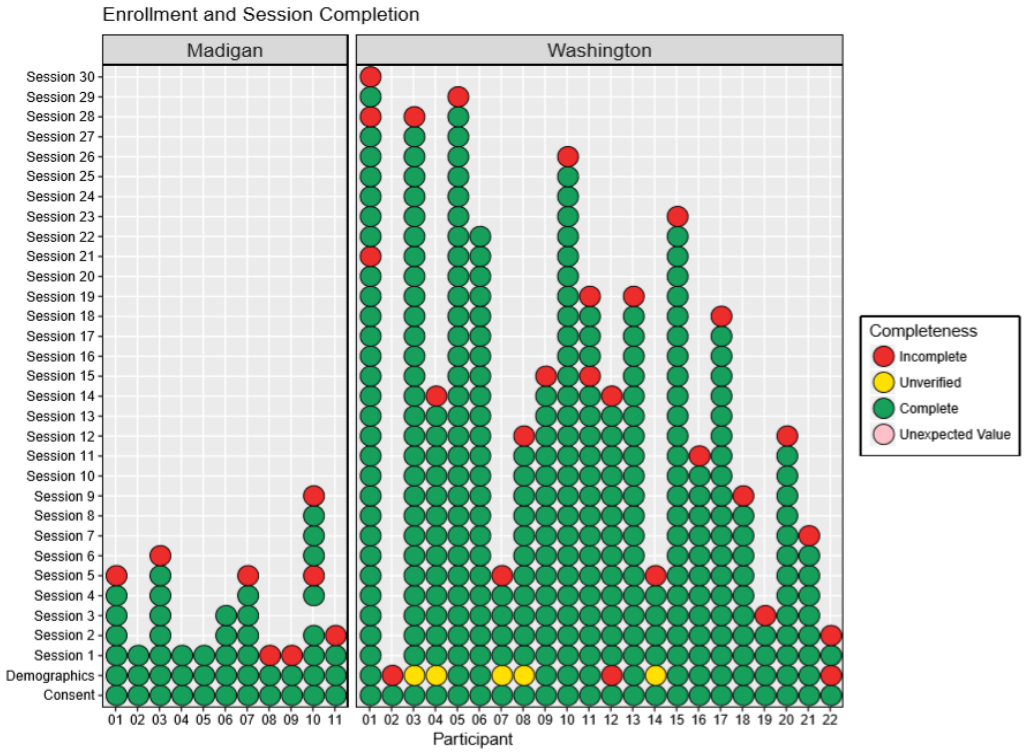
- Site Visits
  - MAMC
    - In August 2017, we conducted a site visit to test the Si dVLogger and to capture data from 2 surgeries in preparation for Aim 2.
    - In April 2018, the UW study coordinator conducted a site visit to run through the data capture process on the newly acquired Xi dVLogger and to test a mobile device that assures the system is operating properly.
    - Three site visits were conducted in June and July of this year (June 28, July 13 and July 20) to train the new site coordinator on use of database and to test the dVLogger system and capture cases.
  - UW/VA
    - In April 2018, (April 12-14), Tim Kowalewski and Jason Kelly spent 3 days at UW to troubleshoot data capture issues.
    - FORT DETRICK, MARYLAND Meeting
      - Dr. Lendvay presented grant update to the DoD on May 10, 2018.

YEAR 4 (July 2018 – July 2019) No Cost Extension – continuation of year 3

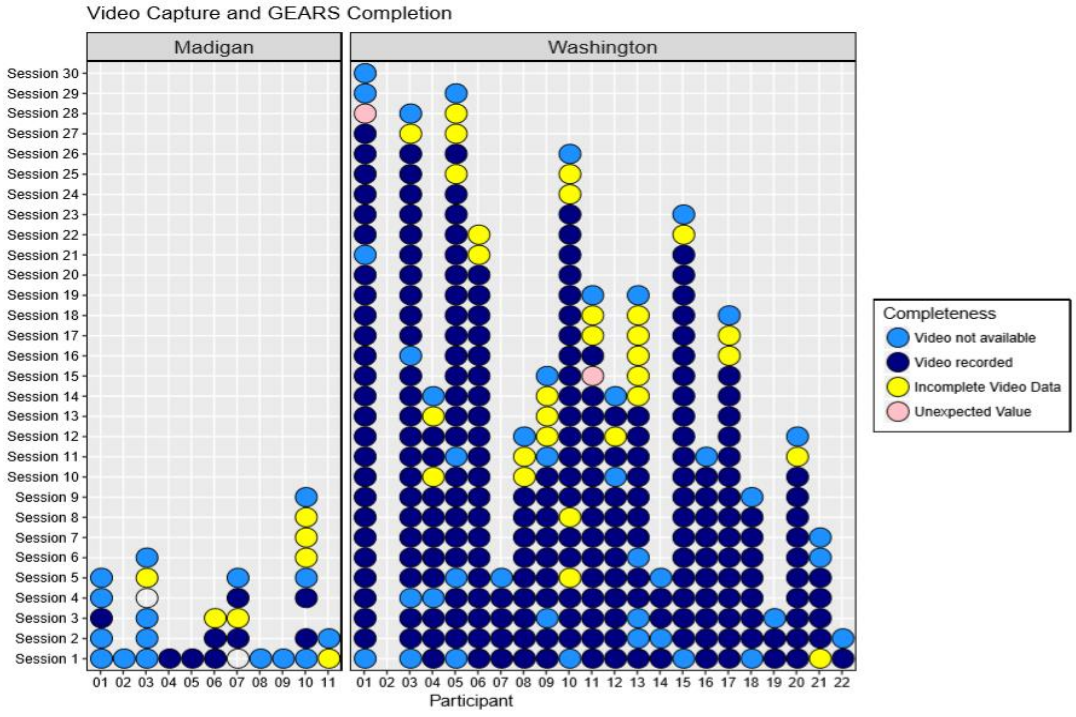
- DoD approved extension from July ’18 to Jan ’20
- Aim 2 Intra-operative RSR subject recruitment and hypothesis testing continues.
- Independent video review of operative performances continues using GEARS.
    - Work Flow
        - Recruited subjects sign consent and complete the intake demographics questionnaire. Subject details are kept at each site and only de-identified data is collected by the team at Minnesota (UMN). All subjects are given a unique identifier based on their location.
        - REDCap is used at all sites to collect data and to randomize the subjects.
        - The data is synchronizing with our central database as designed. UNM continued to optimize REDCap data, conducts dVLogger data integrity checks to ensure all logs are complete and no data is lost and where necessary, investigates incomplete logs and proposes fixes.
        - Data entry was ongoing and REDCap forms were updated as needed for optimal data capture.
    - Aim 2 Recruitment (see tables below)
        - At this point a total of 33 surgeons have consented – 22 at UW/VA and 11 at MAMC.
        - At UW/VA, 3 surgeons are withdrawn; 2 are no longer based at this site,1 surgeon was deemed ineligible after consent. Nineteen surgeons remain actively participating.
        - At MAMC, 2 surgeons deployed during the study and 3 are no longer based at this site. Six remain actively participating.
        - UW HSD Modification #4, was submitted and approved May 3, 2019, allowing surgeons to continue past 24 cases. Four surgeons at UW have signed this updated consent to continue.
    - Table 3: Recruitment data summary by site, n(%)

	Madigan	Washington	Total
Target Enrollment	6	20	26
Randomized Sessions	**34**	**323**	**357**
Completed Sessions	**26**	**302**	**328**
Consented*	11 (183)	22 (110)	33 (127)
Gender			
Female	1 (9)	8 (36)	9 (27)
Male	10 (91)	11 (50)	21 (64)
Other	0 (0)	0 (0)	0 (0)
Specialty			
Urology	2 (18)	8 (36)	10 (30)
Gynocology	0 (0)	9 (41)	9 (27)
General	6 (55)	1 (5)	7 (21)
Thoracic	0 (0)	0 (0)	0 (0)
Cardiothoracic	1 (9)	0 (0)	1 (3)
Handedness			
Left	0 (0)	0 (0)	0 (0)
Right	11 (100)	19 (86)	30 (91)
Ambidextrous	0 (0)	0 (0)	0 (0)
Musical Experience	3 (27)	6 (27)	9 (27)
Deployed	9 (82)	0 (0)	9 (27)

- Figure 6: Aim 2 Enrollment and session completion

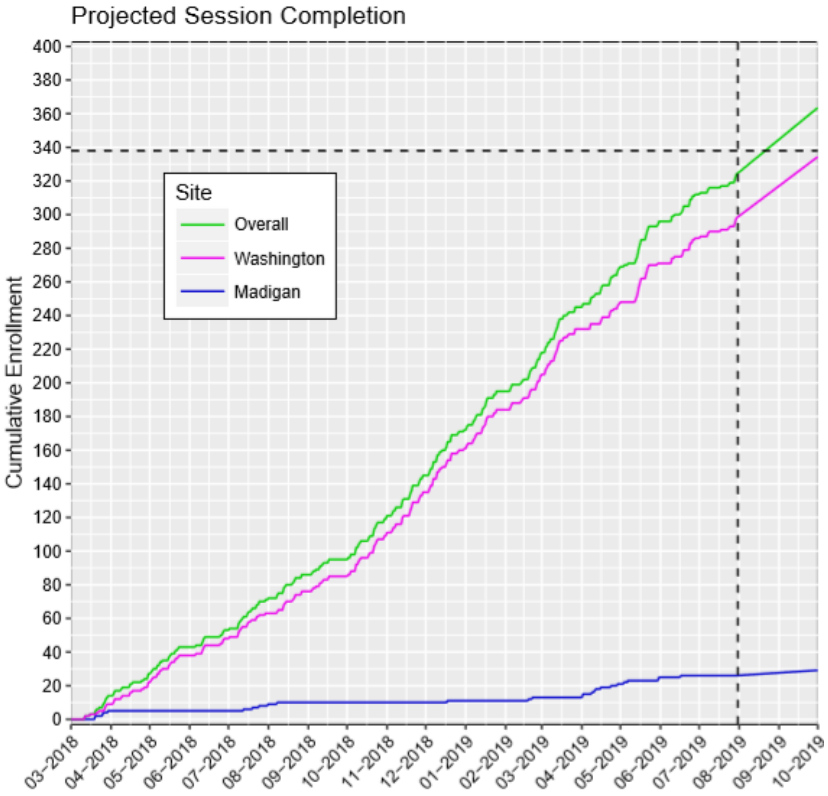


- Figure 7: Video capture and Gears completion

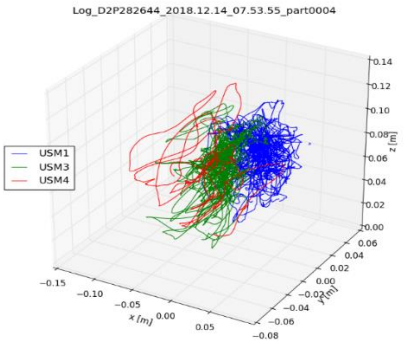


- The study is projected to reach completed sessions by the end of the data collection period. (See Figure 8)

- Figure 8: Projected Session Completion



- Data Capture
  - Aim 2 went live on March 12, 2018. 328 cases were captured - 302 at UW/VA and 26 at MAMC. A second Xi dVlogger was acquired for MAMC increasing case capture capability.
  - UMN continued to monitor collected data for accuracy and completeness and helped resolve issues as they should arise.
- Data Management
  - A new technique was implemented which allowed us to obtain the tooltip positions of the daVinci robot in “world coordinates” (3D coordinates which can be recognized by a human) using the knowledge that when the endoscope camera is not moving, all of the coordinates are in a world coordinate system frame of reference. By dropping all coordinates while the endoscope is moving (during which the other tools aren’t moving) we could get correct Universal Serial Manipulator coordinates to plot, as shown in Figure 9 below.
    - Figure 9



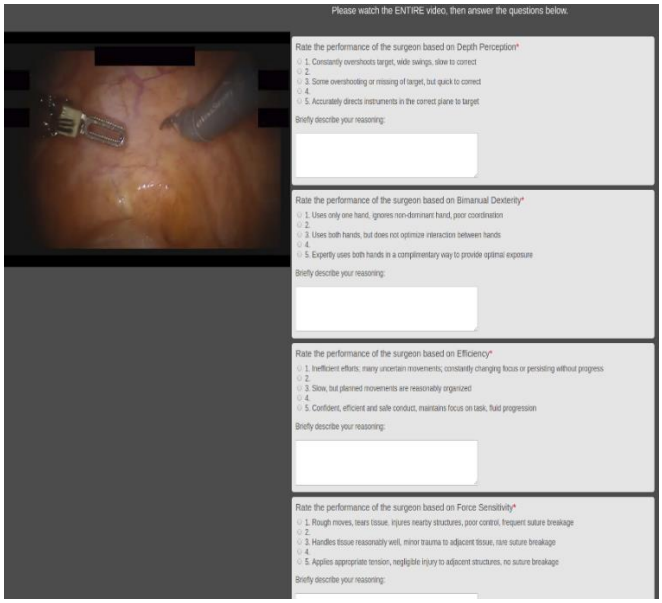
- Video Editing
  - A new website was created using a Google domain name and an apache2 web server, using a LAMP stack (Linux, Apache, MySQL, and PHP). This allowed us to more easily customize how the videos were viewed and created a side-by-side view

of the videos and surveys. PHP scripts and Javascript files were used with CSS to create this.

- Figure 10: Older Zoho site



- Figure 11: Newer site created with an Apache web server:



- Video Review
  - A total of 7 surgeons, with experience performing robotic surgeries, began the video review. A preliminary set of 5 videos were sent to each of these surgeons. After we received ratings for these 5 videos, the initial inter-class correlation coefficient (ICC) was only 0.05. We anticipated 360 videos. Should the expert video review process not lead to better agreement, we sought alternative video review avenues, such as Crowd-sourced Assessment of Technical Skills (CSATS) to complete the review. We initiated discussion with the vendor for a price quote for the review of all videos. Using this rubric for review required permission from the DoD and the UW OSP.
  - For the expert review, we went to a system for properly randomizing the remaining videos so that the correct number of reviewers is obtained for each of the surgical

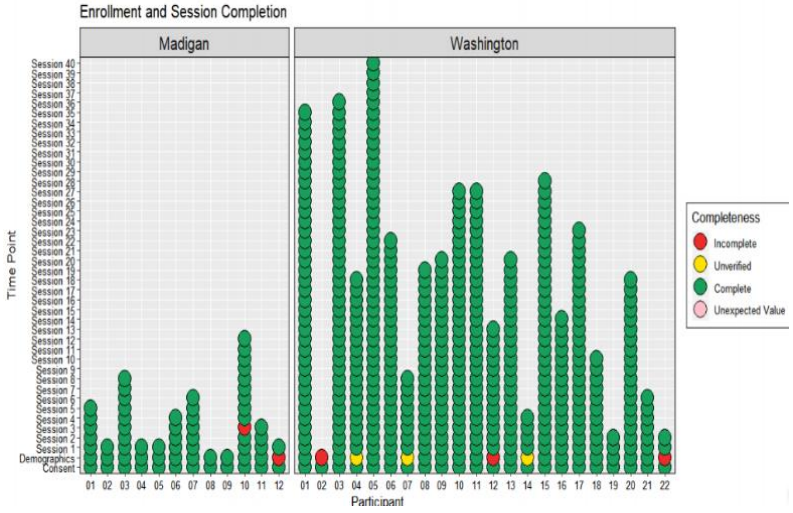
videos.
        - Several sessions were recorded in the last three months of this year, with the total data recorded increasing to 2.8 terabytes of total video and kinematic data.
    - Site Visits
        - The UW Study Coordinator made several trips to MAMC over the course of the year to train staff, troubleshoot data collection issues and capture sessions.

YEAR 5 (July 2019-July 2020) No Cost Extension – continuation of year 3

- Recruitment was completed with 34 consented surgeons – 22 at UW/VA and 12 at MAMC
    - Table 4: Final Aim 2 Recruitment data summary by site, n (%)

	Madigan	Washington	Total
Sessions with Video/Kinematic	17	326	343
Completed Sessions	41	391	432
Consented	12 (200)	22 (110)	34 (131)
Deployed Currently	0 (0)	0 (0)	0 (0)
Lost to follow-up	0 (0)	0 (0)	0 (0)
Gender			
Female	1 (8)	8 (36)	9 (26)
Male	10 (83)	11 (50)	21 (62)
Other	0 (0)	0 (0)	0 (0)
Specialty			
Urology	2 (17)	8 (36)	10 (29)
Gynocology	2 (17)	10 (45)	12 (35)
General	6 (50)	1 (5)	7 (21)
Thoracic	0 (0)	0 (0)	0 (0)
Cardiothoracic	1 (8)	0 (0)	1 (3)
Handedness			
Left	0 (0)	0 (0)	0 (0)
Right	11 (92)	19 (86)	30 (88)
Ambidextrous	0 (0)	0 (0)	0 (0)
Musical Experience	3 (25)	6 (27)	9 (26)
Deployed Previously	9 (75)	0 (0)	9 (26)

- Data capture was completed. We captured a total of 432 Aim 2 cases - 391 at UW/VA and 41 at MAMC. We reached our goal with 343 cases that have complete video/kinematic data.
    - Figure 12: Final Enrollment and session completion

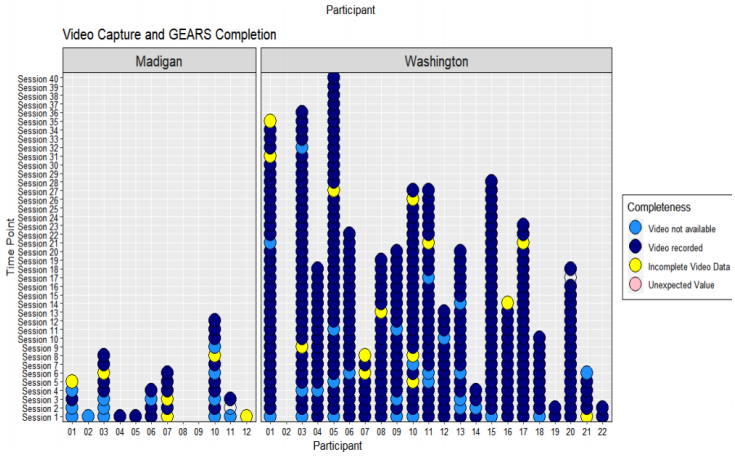


- Data Management
    - We completed the data collection phase, increasing our database by over 80 surgical cases since October 1st, not adjusting for missing kinematic or video data.

- Since October 1, 2019 we collected a total of 238.4 Gigabytes of surgical data and, between October-December, we increased our surgical database by ~19%.
  - Table 5: Cases and Amount of Data

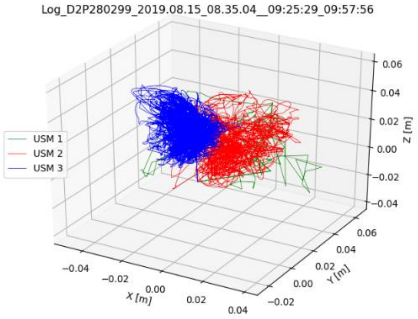
	# Usable Cases (10/01-12/31)	Amount of Data (GB)
University of Washington	47	223.53
Madigan Army Medical Center	7	14.87

- We recorded, uploaded, and backed up a total of 317.8 gigabytes of both video and tooltip kinematic data from these cases.
  - Figure 13: Video capture and gears completion



- Several performance metrics were calculated during the criterion portion of the kinematic data from surgeries recorded, including rate of orientation change (ROC), spectral arc length (SAL), normalized angular displacement (NAD), as well as overall path length, minimum velocity, maximum velocity, and mean velocity.
- Performance metrics from the daVinci's backend API were also calculated from the number of times certain 'events were taking place in the surgery. These included:
  - Number of overall Clutch presses
  - Number of times head was placed in or out of the console
  - Number of times the camera movement buttons were used
  - Number of times energy control was activated to move arms
  - Number of times the finger or foot clutches were pressed
  - Number of times the assignment of specific arms were changed during surgery
  - The amount of time each arm was idle or moving
- Kinematic data were used to conduct exploratory data analysis examining the distributions of individual measures and correlations among measures. Specifics of analytic strategies for evaluating primary and secondary outcomes were discussed.
- Ongoing transformation of the kinematic data into decipherable 3D-coordinate frames is occurring, with an example plot of the 3 surgical arms from a batch of shown below.

- Figure 14: Video editing



  - We enlisted seven expert surgeons to evaluate videos. They were in various stages of the process with anywhere from 7-40 videos per surgeon evaluated.
    - Table 6: Number of videos evaluated by surgeon

Surgeon	Videos Evaluated
Surgeon A	7
Surgeon B	13
Surgeon C	30
Surgeon D	35
Surgeon E	35
Surgeon F	40
Surgeon G	40

  - We recorded expert surgeon evaluations for the past several months using a web domain created to host the robotic surgery videos, which were previously edited to show the first 15 minutes of the surgical performance of the surgeon being studied.
  - The expert surgeons are additionally instructed to review the entire video and rate the video in terms of the GEARS assessment for robotic surgical skill, consisting of: Depth Perception, Bimanual Dexterity, Efficiency, Force Sensitivity, and Robotic Control
  - The biostatistician at University of Washington evaluated the intraclass correlation (ICC) of these surgeon ratings.
  - A Python script was used to randomly assign videos to each surgeon, to avoid any ordering bias.
- Video Review
  - Reliability of expert ratings were evaluated using Krippendorff's alpha. In order to minimize the effects of differences between raters on the outcome of GEARS score and better characterize this outcome, multiple raters will rate each surgery. As specified in the statement of work, linear mixed effects models were used to evaluate the effect of the warm-up procedure on GEARS score. Random intercepts for each rater and each surgeon were incorporated into a linear mixed effects model with a single GEARS score for a specific surgery by a single rater as the outcome.
  - A preliminary set of 40 videos had been sent to each of these evaluating surgeons. The intra-class correlation coefficient (ICC) for cumulative scores among all raters was only 0.05 (95% CI from -0.083 to 0.516). After removing the two raters who rated the fewest videos (7 and 13 videos respectively), leaving five raters who rated 25 to 40 videos each, ICC

improved to a still unacceptably low 0.461 (95% CI from 0.282 to 0.660). The expert video review process had not led to acceptable agreement, thus we sought an alternative video review avenue, Crowd-sourced Assessment of Technical Skills (CSATS), to complete the review.
    - Preliminary analysis of the kinematic and GEARS data showed no intervention effect compared to control.
- Abstract and manuscript drafting completed.
    - Manuscript submitted to the British Journal of Urology

What opportunities for training and professional development has the project provided?

All participants of both Aim 1 and Aim 2 experienced simulation education on the da Vinci robot which may have enhanced their technical skills and expanded their understanding of surgical warm-up. Although our final data did not show a warm-up benefit among the entire cohort of Aim 2 subjects, some of the subjects anecdotally reported that spending a few minutes on the VR simulator prior to doing their cases help hone their attention and did, in fact, make them feel more prepared for the surgery. Our study did not test the cognitive aspects of warm-up benefit and this may be an additional benefit for our subjects during the study.

How were the results disseminated to communities of interest?

See Peer-reviewed Manuscript and Conference Presentations list

What do you plan to do during the next reporting period to accomplish the goals?

Nothing to report

IMPACT:
What was the impact on the development of the principal discipline(s) of the project?

Contradictory to a number of smaller, less rigorous studies on the role of warm-up in the operating room, our study did not show a benefit in technical performance. This, in it of itself, is a novel finding and may suggest that 1) our warm-up curriculum did not prime surgeons enough, 2) aspects of surgical skill not tested during the study were impacted but not captured with our metrics, and/or 3) warm-up is helpful for steps of a surgery not observed in this study. If anything, our results prompt us to consider other ways in which some type of priming could help surgeons optimize performance. Our data should prompt other investigators to look for additional performance metrics to determine warm-up value.

What was the impact on other disciplines?

Within the entire surgical discipline, our study was inclusive and can generalize between surgical subspecialties. Our results may also be relevant for other professional technical skills people try to measure and perhaps apply a warm-up in front of. The template for what types of technical performance metrics, the use of crowdsourcing to assess skill, the use of automated performance metrics (APMs) to assess skill can be used across many different technical skills assessment disciplines. In addition, the model of using subjects to act as their own controls, we believe was a way to reduce confounding when subjects are just randomized to one or the other treatment/control technical skills assessment arm. We learned the challenges of doing group randomization during our previous study.

What was the impact on technology transfer?

There were no inventions or start-ups resulting from this work.

What was the impact on society beyond science and technology?

There were behavior changes we observed among surgeons in that surgeons who performed their routine day-to-day surgeries, in some cases, are choosing to continue a brief warm-up prior to their surgeries – outside of the study.

CHANGES/PROBLEMS:

Year 1: There was nothing to report

Year 2: There was nothing to report

Year 3: Added capability of capturing cases on both Si and Xi platforms to allow opportunity to increase the rate of case capture to meet study goals.

Year 4: Added capability of capturing cases on both Si and Xi platforms to allow opportunity to increase the rate of case capture to meet study goals.

Changes in approach and reasons for change
Actual or anticipated problems or delays and actions or plans to resolve them

Year 1: Contractual and Funding Delays

- Delays in IRB approvals at VA
- Delays in funding at several sites
    - Due to the nature of funding a military site and the need for all funds to be used within that fiscal year we had to bring back the money allocated to MAMC and fund any manpower hours required centrally.
    - Funding contract with Florida Hospital took longer than anticipated but was able to be worked out. Specifically, the University of Washington (a state-funded institution) was prohibited from providing a sub-contract to an institution that did not have an Equal Opportunity Employment (EOE) clause (Florida Hospital). This was resolved and the sub-contract was granted.

Year 2: Enrollment Barriers

- While one site (MAMC) had proven exceptional, enrollment at UW+VA and Florida were lower than projected. The principal barrier appeared to be subject inconvenience. Slower enrollment at UW/VA sites were due to staffing issues which have been resolved by the addition of dedicated staff MAMC's simulator was in a room beside the clinic, near the physicians' offices, and was used in surgery one to two days per week. The subjects were in a relatively private space, with nearly unlimited access to the equipment, and were not required to access the operating suites or change into scrubs.
    - In comparison, the UWMC simulator was located in an operating room (OR), and is not available during surgery hours (i.e., ~07:00-19:00 weekdays). All users are required to change into scrubs.
    - The VA's simulator, when not in use, was packed in a storage closet adjacent to the OR; access to the equipment was dependent on the OR's usage. All users were required to

change into scrubs. Further, the General Surgery residents shared a key-card, limiting access to the locker rooms and OR.
    - In addition, the General Surgery residents in the UW/VA program were required to work a minimum of 80 hours per week. The General Surgery Director had determined that time related to this study may not be included in that commitment, and must be scheduled separately. This restriction had been a deterrent to proceeding through the study.
- At FL Hospital, the simulator facility was approximately 20 minutes' drive from the hospital proper. Subjects were understandably reluctant to invest in the additional commute.
- Randomization of consented subjects had been impacted by clinical rotations and busy schedules, limiting residents to participate once they had consented to the study.

Year 3: Enrollment Barriers

- Florida Hospital dropped out so volume was increased at the UW/VA and MAMC sites to offset that
- Loss of the MAMC site coordinator in April was reconciled by the addition of a new site coordinator who completed training in July.

Year 4: Enrollment barriers

- Low volume surgeons were reaching target for per surgeon case volume.
    - The IRB approved expanding the case volume targets for other surgeons
- UW project manager went to MAMC to capture video

Changes that had a significant impact on expenditures

The biggest impact was that in the first year, UW and our Florida site were unable to establish a contracting agreement because of two differing Equal Opportunity Employment clauses between the institutions. This added an entire year of delay in subject recruitment because IRBs could not be reconciled until the contract could be reconciled. Furthermore, Florida ultimately was unable to execute on the Aim 2 work and much of the Aim 1 work and so the delay in subject recruitment drew out the spending of the budget. Thus, we received two No Cost Extensions.

Year 1: There was nothing to report

Year 2: Since Florida has not been as successful at recruitment and retention, more resources were needed at University of Washington/VA to cover the additional FTE for Year 3. The DV Logger was able to extract the necessary kinematic data from the daVinci Robot, and thus we sought approval from the DoD to repurpose monies allocated to the optical tracker to personnel FTE, especially at UW and UMN. A formal budget modification request was submitted in Year 3.

Year 3: The delay getting started in year 1 led to a slower burn rate for expenditures. This led to a budget surplus that allowed for financial support of the 18 month NCE which was approved July 14, 2018. The DoD approved a formal budget modification repurposing monies allocated to the optical tracker to personnel FTE, especially at UW and UMN and using dVLoggers to capture kinematic and video data.

Year 4: The delay getting started in year 1 led to a slower burn rate for expenditures. This led to a budget surplus that allowed for financial support of the 18 month NCE which was approved July 14, 2018.

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

Nothing to report

Significant changes in use or care of human subjects

Nothing to report

Nothing to report

5. PRODUCTS: Publications, conference papers, and presentations

Journal publications.

French A, Lendvay TS, Sweet RM, Kowalewski TM. “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. 2017 Jul 1;12(7):1161-70.

Dockter RL, Lendvay TS, Sweet RM, Kowalewski TM. “The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data.” International journal of computer assisted radiology and surgery. 2017 Jul 1;12(7):1151-9.

French A, Kowalewski TM. “Laparoscopic skill classification using the two-third power law and the isogony principle”. In Frontiers in Biomedical Devices 2017 Apr 10 (Vol. 40672, p. V001T12A001). American Society of Mechanical Engineers.

Kelly J, Petersen A, Lendvay TS, Kowalewski TM. “The Effect of Video Playback Speed on Surgeon Technical Skill Perception”, Int J CARS 15, 739-747 (2020).

Kelly JD, Petersen A, Lendvay TS, Kowalewski TM, “Bidirectional Long Short-Term Memory for Surgical Skill Classification of Temporally Segmented Tasks”, Int J CARS , [in review; CARS-D-20-00212R1]

Kelly JD, Heller N, Petersen A, Lendvay TS, Kowalewski TM, “The Effect of Video Playback Speed on Perception of Technical Skill in Robotic Surgery”, Int J CARS, [in review; CARS-D-20-00299R1]

Kelly JD, Nash M, Heller N, Lendvay TS, Kowalewski TM, “Temporal Variability of Surgical Technical Skill Perception in Real Robotic Surgery”, Int J CARS, [in review; CARS-D-00322R1

Kelly JD, Kowalewski TM, Brand T, French A, Nash M, Meryman L, Heller N, Organ N, George E, Smith R, Sorensen MD, Comstock B, Lendvay TS, “Robot-Assisted Surgery Readiness: Virtual Reality Warm-Up
Prior to Robot-Assisted Surgery”, British Journal of Urology, [Submitted; BJU-2020-1201]

Books or other non-periodical, one-time publications.

Nothing to Report

Other publications, conference papers and presentations.

"Predicting Surgical Skill from the First N Seconds" at the IPCAI conference, Barcelona, Spain June 2017

Dr. Lendvay presented grant update to the DoD on May 10, 2018

Abstract for podium at the Society of Pediatric Urology meeting Sept 14-16, 2018, Atlanta

Aim 1 results were also presented at Design of Medical Devices Conference, University of Minnesota, April 16 -18, 2019

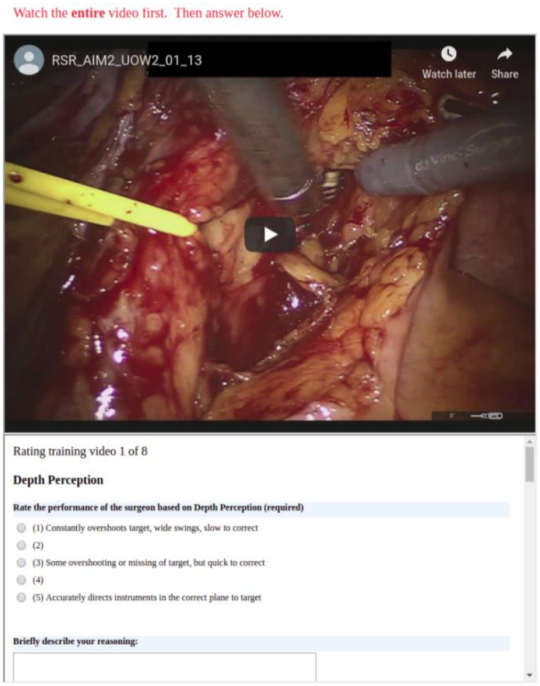
Kelly, Jason "Analysis and Control of Temporal Biases in Surgical Skill Evaluation" PhD diss., University of Minnesota, 2020.

Website(s) or other Internet site(s)

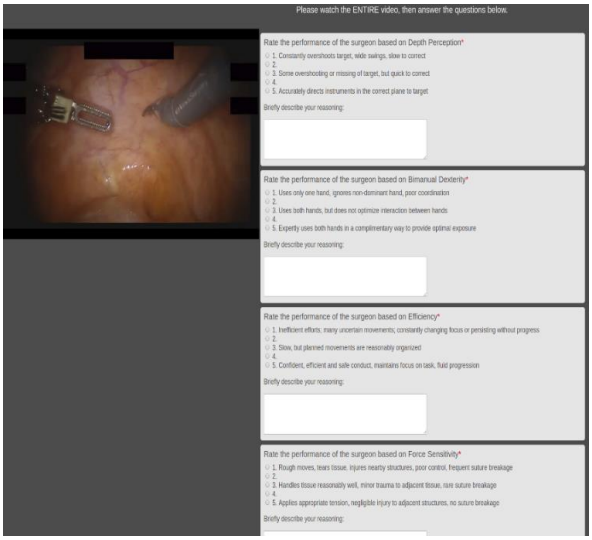
Nothing to report

Technologies or techniques

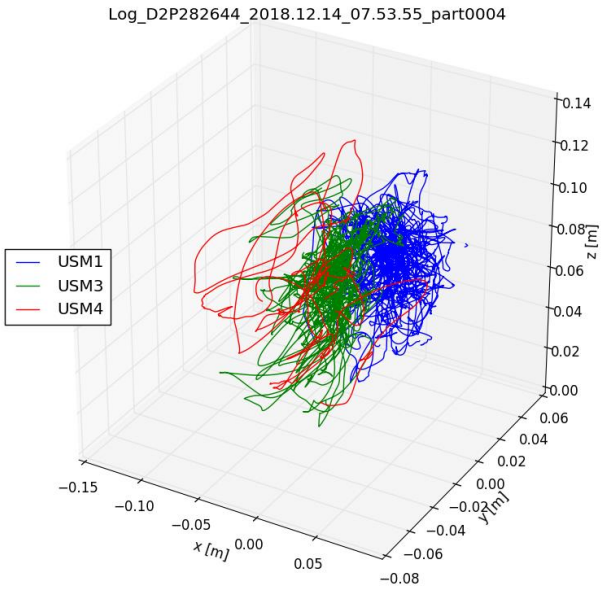
A new website was created using a Google domain name and an apache2 web server, using a LAMP stack (Linux, Apache, MySQL, and PHP). This allows us to more easily customize how the videos are viewed and create a side-by-side view of the videos and surveys. PHP scripts and Javascript files were used with CSS to create this.



The newer site created with an Apache web server



A new technique has been implemented which allows us to obtain the tooltip positions of the daVinci robot in “world coordinates” (3D coordinates which can be recognized by a human) using the knowledge that when the endoscope camera is not moving, all of the coordinates are in a world coordinate system frame of reference. By dropping all coordinates while the endoscope is moving (during which the other tools aren’t moving) we can get correct Universal Serial Manipulator coordinates to plot, as shown below.



UMN reworked software to provide high definition video acquisition and compression and successfully installed and monitored the data acquisition from all sites. The data synchronizes with our central database as designed. UNM rectifies the Google Drive videos, REDCap data, and dVLogger data ensuring all logs are complete and no data is lost and where necessary, investigates incomplete logs and proposes fixes. This method of reliable video capture, data tagging, and storage has a universal application in any training and skills assessment program

Inventions, patent applications, and/or licenses

Nothing to report

Other Products

Nothing to report

PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name: Dr. Thomas Lendvay
Name: Jason Kelly
Name: Dr. Timothy Kowalewski

Name: Lois Meryman

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

No Changes

What other organizations were involved as partners?

University of Washington, Seattle, WA, – *Facilities, Collaboration*

VA Puget Sound Health Care System, Seattle, WA - *Facilities*

Madigan Army Medical Center, Henry M. Jackson Foundation, Tacoma, WA, - *Facilities, Collaboration*

University of Minnesota, Minneapolis, MN, - *Collaboration*

Florida Hospital Nicholson Center, Orlando, FL, - *Facilities, Collaboration*

SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS

Nothing to report

QUAD CHARTS: *If applicable, the Quad Chart (available on https://www.usamraa.army.mil) should be updated and submitted with attachments.*

6. APPENDICES:

Peer reviewed manuscripts

Appendix A
French A, Lendvay TS, Sweet RM, Kowalewski TM. "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. 2017 Jul 1;12(7):1161-70.

Appendix B
Dockter RL, Lendvay TS, Sweet RM, Kowalewski TM. "The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data". International journal of computer assisted radiology and surgery. 2017 Jul 1;12(7):1151-9.

Appendix C
French A, Kowalewski TM. "Laparoscopic skill classification using the two-third power law and the isogony principle". In Frontiers in Biomedical Devices 2017 Apr 10 (Vol. 40672, p. V001T12A001). American Society of Mechanical Engineers.

Appendix D
Kelly J, Petersen A, Lendvay TS, Kowalewski TM. "The Effect of Video Playback Speed on Surgeon Technical Skill Perception", Int J CARS 15, 739-747 (2020).

Appendix E
Kelly JD, Petersen A, Lendvay TS, Kowalewski TM, "Bidirectional Long Short-Term Memory for Surgical Skill Classification of Temporally Segmented Tasks", Int J CARS , [in review; CARS-D-20-00212R1]

Appendix F

Kelly JD, Heller N, Petersen A, Lendvay TS, Kowalewski TM, “The Effect of Video Playback Speed on Perception of Technical Skill in Robotic Surgery”, Int J CARS, [in review; CARS-D-20-00299R1]

Appendix G
Kelly JD, Nash M, Heller N, Lendvay TS, Kowalewski TM, “Temporal Variability of Surgical Technical Skill Perception in Real Robotic Surgery”, Int J CARS, [in review; CARS-D-00322R1]

Appendix H
Kelly JD, Kowalewski TM, Brand T, French A, Nash M, Meryman L, Heller N, Organ N, George E, Smith R, Sorensen MD, Comstock B, Lendvay TS, “Robot-Assisted Surgery Readiness: Virtual Reality Warm-Up Prior to Robot-Assisted Surgery”, British Journal of Urology, [Submitted; BJU-2020-1201]

7. Conference presentations

Appendix I
"Predicting Surgical Skill from the First N Seconds" at the IPCAI conference, Barcelona, Spain June 2017

Appendix J
Abstract for podium at the Society of Pediatric Urology meeting Sept 14-16, 2018, Atlanta

Appendix K
Aim 1 results were also presented at Design of Medical Devices Conference, University of Minnesota, April 16 -18, 2019

Appendix L
Kelly, Jason "Analysis and Control of Temporal Biases in Surgical Skill Evaluation" PhD diss., University of Minnesota, 2020.

Appendix A

Int J CARS manuscript No.
(will be inserted by the editor)

Predicting Surgical Skill from the First N Seconds of a Task

Value over Task Time Using the Isogony Principle

Anna French · Thomas S. Lendvay M.D. · Robert M. Sweet M.D. · Timothy M. Kowalewski Ph.D.

Received: date / Accepted: date

Abstract *Purpose:* Prior attempts at surgical skill evaluation have focused predominantly on diagnosis using task-specific maneuvers. These maneuvers required surgical expertise to identify and are observed over the course of a full task. The aim of this investigation is to propose features for automated skill evaluation that are relevant regardless of the surgical training task the tools perform. A secondary goal is to diagnose skill without requiring the complete time series of data from a given trial.
Methods: Features calculated from the isogony principle are used to train four common machine learning algorithms from dry-lab laparoscopic data gathered from three common training exercises. These models are used to predict the binary or ternary skill level of a surgeon. K-fold and leave-one-user-out cross-validation are used to assess the accuracy of the generated models.
Results: It is shown that the proposed scalar features can be trained to create 2-class and 3-class classification models that map to Fundamentals of Laparoscopic Surgery (FLS) skill level with median 85% and 63% accuracy in cross validation, respectively, for the targeted dataset. Also, it is shown that the 2-class models can discern class at 90% of best-case mean accuracy with only 8s of data from the start of the task.
Conclusion: Novice and expert skill levels of unobserved trials can be discerned using a state vector machine trained with parameters based on the isogony principle.

Anna French
Department of Mechanical Engineering University of Minnesota, Minneapolis, MN
E-mail: afrench@umn.edu

T. Lendvay
Department of Urology, Seattle Children's Hospital, Seattle, WA

R. Sweet
Department of Urology, University of Washington, Seattle, WA

T. Kowalewski
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN

Seattle Children's — Internal

The accuracy of this classification comes within 90% of the classification accuracy from observing the full trial within 10 seconds of task initiation on average.

Keywords Surgical Skill Evaluation · Computer aided decision · Tracking systems

1 Introduction

Surgical technical skill directly impacts patient health outcomes, as shown in [2]. An accurate automated surgical skill evaluation system would consequently be an important tool in reducing a surgical patient's injury risk. A system able to deliver evaluations immediately following a training module would also prove beneficial to surgeons in training, since formative feedback is a lauded goal in surgical training [12]. Knowing when one makes an error is invaluable information that molds good behavior. It is a core educational principle that spans disciplines [13]. Current methods introduce time barriers to feedback, either by requiring humans to deliver ratings or by using scoring models which require the compilation of scored trial data each time a new training technique is developed. To accomplish timely feedback, a scoring method must be developed that depends neither on human intervention nor on prior probability distributions for features specific to a particular task.

Past attempts at developing an automated skill evaluation system have focused on diagnosing skill using task-specific performance measurements. In [10], a method using linear discriminant analysis and tool motion features achieved accuracy in the 90% range, however the assessment method was tailored to four-throw suturing tasks and defining task features (surgemes) required a surgeon's expertise. Investigation from [1] reports similar classification success but also used features very specific to septoplasty, and classified by segmenting and analyzing the stroking motions of the cottle. The investigation in [8] showed the crowd is capable of discerning surgical skill concordant with the current gold standard (an expert panel) and can generate a skill evaluation weeks faster than the expert panel. However, these scores still depend on human intervention and also introduce several hours of lag time between task completion and score delivery to the trainee. A method that does not require task completion to diagnose skill and can model skill regardless of which task the surgeon performs would be preferred.

The Fundamentals of Laparoscopic Surgery is a high-stakes certification exam of cognitive and technical laparoscopic skills [4, 5, 11]. It is now often required for graduation and or board certification among laparoscopic curricula. It has been extensively validated and even shown to correlate with patient outcomes [14]. However, there are some limitations. The score is based on task time and penalty counts (e.g. dropping an item, cutting outside a boundary, loose knots). [7] found that the weights used in computing FLS score greatly emphasize task time, rendering the penalties virtually irrelevant. This suggests that FLS score provides little or no practical value over task time. However the value of FLS scoring (or, implicitly, FLS task time) in its link to patient outcomes remains undisputed and therefore valuable as a measure of surgical technical skill. Our observations in operating rooms, surgical simulator sessions, and among trainees reveal that a subject's approximate level of technical skill is often evident very

Seattle Children's — Internal

quickly–within tens of seconds–when watching video of their tool motions during a procedure. This suggests that skill evaluation can be correctly approximated with a fraction of the time it takes to do a procedure–be it an FLS task or a surgical procedure. This would alleviate the need for trainees and proctors to wait until the end of a task to receive a skill evaluation. However, it is unclear either what motion features to identify or how much time is needed before obtaining adequate confidence in such an assessment, i.e., how many seconds are required to predict an FLS score?

1.1 Isogony Principle

The isogony principle may provide some value to tool motion-based skill evaluation. In [9], subjects were recorded drawing shapes of various curvature with the goal of relating curvature of the drawn shape to speed of the pen tip. A relation between these two parameters was determined using the isogony principle as:

$$v(t) = \gamma k(t)^{1/3}$$

where $v(t)$ is the instantaneous velocity of the tip of the pen, $k(t)$ is the local curvature that the tip of the pencil traces, and γ is the velocity gain factor parameter relating $v(t)$ and $k(t)$. In [9], it was asserted that $v(t)$ can be predicted from $k(t)$ based on a constant value of γ for a given segment of motion.

This investigation extends the velocity gain factor relationship to 3D tool motion, using the velocity and curvature from the 3D space. For the purpose of this study, we do not assume constant values of γ, and instead choose to observe the behavior of the γ parameter:

$$\gamma(t) = \frac{v(t)}{k(t)^{1/3}}$$

For the $k(t)$ parameter, the radius of curvature was used:

$$k(t) = \frac{(1 + v(t)^2)^{3/2}}{a(t)}$$

1.2 Hypotheses and Objectives

Based on the property from [9], several hypotheses were drawn. First, it was hypothesized that the variability of the γ parameter between novice surgeons will be small. This was drawn from the idea that novice surgeons will adhere more to their "natural" hand motion pattern, while the more practiced motions of experienced surgeons will vary from this natural motion pattern.

Second, it was hypothesized that scalar parameters such as the mean (μ) and standard deviation (σ) of a trial's γ for each hand can be used as features to train machine learning algorithms and coarsely predict the Fundamentals of Laparoscopic Surgery (FLS) score of unobserved trials.

Third, it was hypothesized that the full duration of the task is not required to evaluate skill since $\gamma(t)$ is easily observable at any point in the task. Accordingly,

Seattle Children's — Internal

investigation was made into the minimum number of seconds of data from a trial required to discern the subject's FLS class with an acceptable level of accuracy. This ability would provide significant value over task time related features, which require probability distributions based on the results of previous users, and would not be agile to changes in training. A secondary hypothesis is that prediction accuracy will increase as more time is included, but gains will taper off.

Results from both task-specific and task-blind models were generated and are reported here. Task-specific models are models trained using only samples from a specific task, and are included as a basis for comparison. Task-blind models are models where data across all tasks were included in training, and were used to predict the skill of any task. Both leave-trial-out cross validation and leave-one-user-out (sometimes referred to as leave-surgeon-out, and abbreviated here as LOUO) validation methods were used to test the accuracy of the binary and ternary classification algorithms developed.

There are three key contributions offered by this paper. First, we introduce isogony as a potentially useful feature in surgical skill evaluation. Second, we introduce task invariance as a desired attribute of skill evaluation. Third, we introduce the notion of estimating skill normally evaluated over the course of a full task from a partial task observation, i.e. predicting final scores from N seconds.

2 Methods

2.1 Dataset

This investigation used the dataset established in the Electronic Data Generation and Evaluation (EDGE) study described by [7]. This study gathered video, tool motion and demographic data on 98 different surgeons performing typical FLS tasks. From this data set, 108 peg transfer, 63 suturing and 124 circle cutting tasks were used for this study.

Each instance where data was recorded while a subject was performing a particular FLS task will be referred to as a "trial." Within the dataset, each trial is comprised of a 30Hz fixed camera-position video recording of the laparoscopic tools interacting with the training field, numeric data documenting the position, orientation and grasp force of the tool tips corresponding to each frame in the video, an FLS score ranking the subject's skill level based on their performance in each trial, and demographic information relating relevant information about the subject such as their dominant hand and experience level.

The FLS score alone was used to establish skill groups within each of the three tasks. This resulted in the FLS expert class (any trial with FLS scores above a threshold of OSATS scores from identity-blind review by two faculty surgeons for each task; see [6]); the FLS novices (trials from the bottom 15th-percentile of FLS scores within each task); and FLS Intermediates (trials from the 15th-percentile range about a midpoint between the lowest FLS Expert score and highest FLS Novice score, for each task). The video portion of the data was not used. This choice of criteria gave us complete trials from 67 FLS Novices, 71 FLS Intermediates, and 157 FLS Experts.

Seattle Children's — Internal

2.2 Analysis Methods and Algorithms

The mean and standard deviation of the γ parameter of each trial for the dominant (d) and non-dominant (nd) hands are the features selected for evaluation. These are referred to as $\sigma(\gamma_d), \mu(\gamma_d), \sigma(\gamma_{nd})$ and $\mu(\gamma_{nd})$. These four features were calculated for each trial, and were used along with their FLS class to train several different machine learning algorithms to classify skill level. The accuracies these trained algorithms obtained were used as evidence of feature strength. The algorithms used for testing were logistic regression (LR), support vector machine (SVM), linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA).

The accuracy of each trained model was evaluated using k-fold cross validation (with k=10) and leave-one-user-out validation. For the k-fold, each fold was assigned an equal number of trials from each class. The k-fold cross validation was performed N=10 different times, where a new set of k-folds was selected and evaluated for each iteration of N, which resulted in kN different models trained and evaluated for each machine learning algorithm. Note that these are folds created using each individual trial, hence it is partitioning in a leave-trial-out manner.

Accuracy was also evaluated using leave-one-user-out (LOUO) for all Q surgeons. Each surgeon has r different trials in the database, where r may differ for each surgeon. In this method, each surgeon takes a turn as the test set while the other $Q-1$ are used for the training set to generate the models. The accuracy is reported by evaluating the classification results of each of the r trials for each of the Q surgeons. Feature strength and model accuracy were assessed separately for both 2-class classification (discriminating between novice and expert) and 3-class classification (discriminating between novice, intermediate, and expert).

Models were generated in both task-specific and task blind manners. Task-specific models were trained using only trials where a specific task was performed, and their accuracy was tested using only trials from that specific task. Task-blind models were trained using all trials regardless of task and were used to create predictions of any trials regardless of class. The accuracy of task-blind model predictions for each specific class was also analyzed, where the model was trained task-blind but the testing set was partitioned to analyze how well the task-blind model can predict the skill for each specific task.

The minimum period of time required for acceptable prediction accuracy was evaluated by taking successively longer series of time from the beginning of each trial to time t and calculating $\mu({}^t\gamma_d)$, $\sigma({}^t\gamma_d)$, $\mu({}^t\gamma_{nd})$, and $\sigma({}^t\gamma_{nd})$ based on those different time periods. The feature $\mu({}^t\gamma_d)$ is the feature $\mu(\gamma_d)$ calculated from the γ values from the start of the training exercise until time t, where $n(t)$ represents the number of time-steps included in the range [0,t] (data was recorded at 30Hz, so $n = 30t$):

$$\mu({}^t\gamma_d) = \frac{1}{n(t)} \sum_{i=0}^{n(t)} \frac{v_d(i)}{k_d(i)^{1/3}} \qquad \sigma({}^t\gamma_d) = \sqrt{\frac{1}{n(t)} \sum_{i=0}^{n(t)} \left(\frac{v_d(i)}{k_d(i)^{1/3}} - \mu({}^t\gamma_d) \right)}$$

The $\mu({}^t\gamma_{nd})$ and $\sigma({}^t\gamma_{nd})$ are calculated by the same method, but with the non-dominant hand measurements. The above features were generated for each integer-valued time period second within $t = (1, 30)s$, where 30s was chosen since all task times in the data set were greater than this amount. This created 30 different

Seattle Children's — Internal

Table 1: Median and standard deviation of the $\sigma(\gamma_d)$ and $\sigma(\gamma_n d)$ over all subjects.

		$\sigma(\gamma_d)$ med(standard deviation)	$\sigma(\gamma_{nd})$ med(standard deviation)
PegTx	Nov	13.91 (4.22)	11.52 (5.83)
	Exp	22.86 (7.03)	20.98 (8.71)
Cutting	Nov	9.24 (5.83)	9.72 (3.88)
	Exp	18.37 (5.97)	18.83 (10.61)
Suturing	Nov	9.09 (2.31)	9.45 (2.76)
	Exp	14.85 (5.83)	17.66 (7.90)
All Tasks	Nov	10.19 (5.14)	9.93 (4.23)
	Exp	18.80 (7.04)	19.45 (9.57)

groups of $\mu({}^t\gamma_d)$, $\sigma({}^t\gamma_d)$, $\mu({}^t\gamma_{nd})$, and $\sigma({}^t\gamma_{nd})$ specific to the period of time they were calculated from. Each of these 30 groups were then passed through the same machine learning algorithms and validation process as for $\sigma(\gamma_d), \mu(\gamma_d), \sigma(\gamma_{nd})$ and $\mu(\gamma_{nd})$, yielding a mean accuracy μ_t for each group. Trends for the value of μ_t for $t = (1, 30)s$ for each different machine learning algorithm were then plotted, and are displayed in the results section. The minimum t required to get within 90% of the observed settling accuracy is reported in Table 2.

In addition, these methods have been validated against other validated methods for skill classification, such as those in [3]. The validation methods trains three different models using either tool path lengths (PL), economy of motion (EOM) and motion smoothness (MS) as features using SVM.

3 Experimental Results

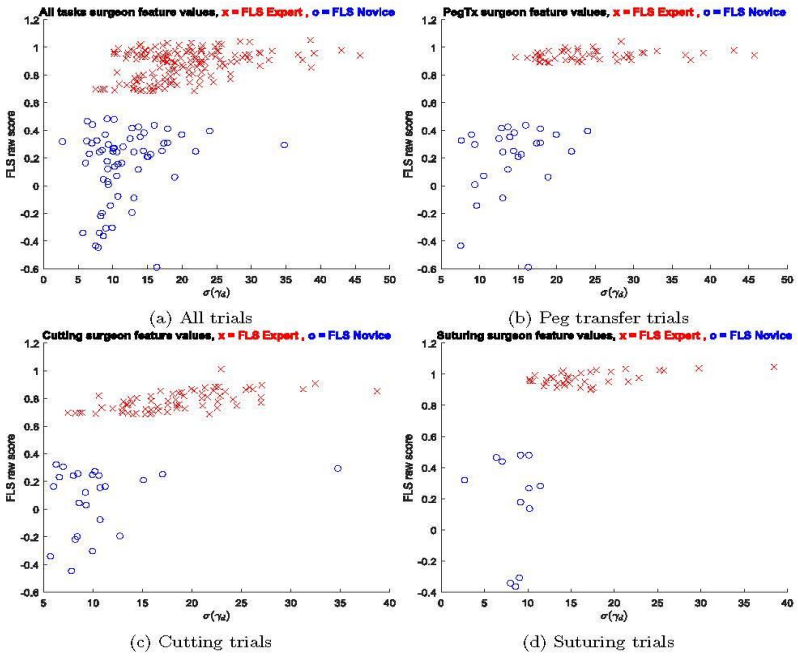
3.1 Two Class Classification: FLS Expert vs. FLS Novice

Figures 1a-1d show the distribution of FLS scores plotted against the $\sigma(\gamma_d)$ feature of each trial for a given subject, where the marker type and color specifies expertise. Recall that $\sigma(\gamma_d)$ and $\sigma(\gamma_{nd})$ are features representing the intra-subject standard deviations (the subject's standard deviation for motion during a given trial). Table 1 details the median and range of the $\sigma(\gamma_d)$ and $\sigma(\gamma_{nd})$ features for each subject. It is observable from here that the inter-subject medians for novices are much lower and have much smaller inter-subject standard deviation.

Figures 2a-2c show six example plots of the γ_D parameter calculated for each time step in the first 20 seconds of a trial.

Figure 3a shows the statistics reporting the mean accuracy of each model trained to classify between FLS novice and FLS expert trials. Five different types of models were generated and tested at each round. The left four box plots represent the model prediction accuracy based on training using a single feature. The rightmost column used all four features to train the model. The mean accuracy of each model trained in each k-cross validation and each N-iteration was recorded and used to generate the box plots. Thus the statistics displayed are values calculated over the kN models generated in the cross validations. These models were trained task-blind, meaning all trials regardless of task type were used to train the model. They were also tested task-blind, meaning the results reported here are the accuracy over all trials regardless of task.

Seattle Children's — Internal



(a) All trials (b) Peg transfer trials

(c) Cutting trials (d) Suturing trials

Fig. 1: FLS score vs $\sigma(\gamma_d)$ feature calculated over the full duration of each trial (dominant hand). Each point represents the feature value for one trial. Raw data displayed for all trials and each trial individually.

For all provided box plots, the 25th and 75th percentiles are the lower and upper box boundaries while the median is the central line. The whiskers extend to the most extreme non-outlier points, and the + are considered outliers. The models trained using single features from each hand have agreement not far from the model trained on all features, which shows a median agreement with desired skill class of 85%.

Figure 4a shows the box plots representing model prediction accuracy using LOUO. These models are trained using all four features and the labeled machinery type. This figure shows a median model prediction accuracy of between 80% and 100% depending on model type. Note that LOUO tests the accuracy of each user individually, i.e. each user has performed n trials and a prediction accuracy is assessed for each individual user based on the percentage of those n trials that were correctly classified. The LOUO box plot displays the crowd tendencies of the percent accuracies of each user.

Seattle Children's — Internal

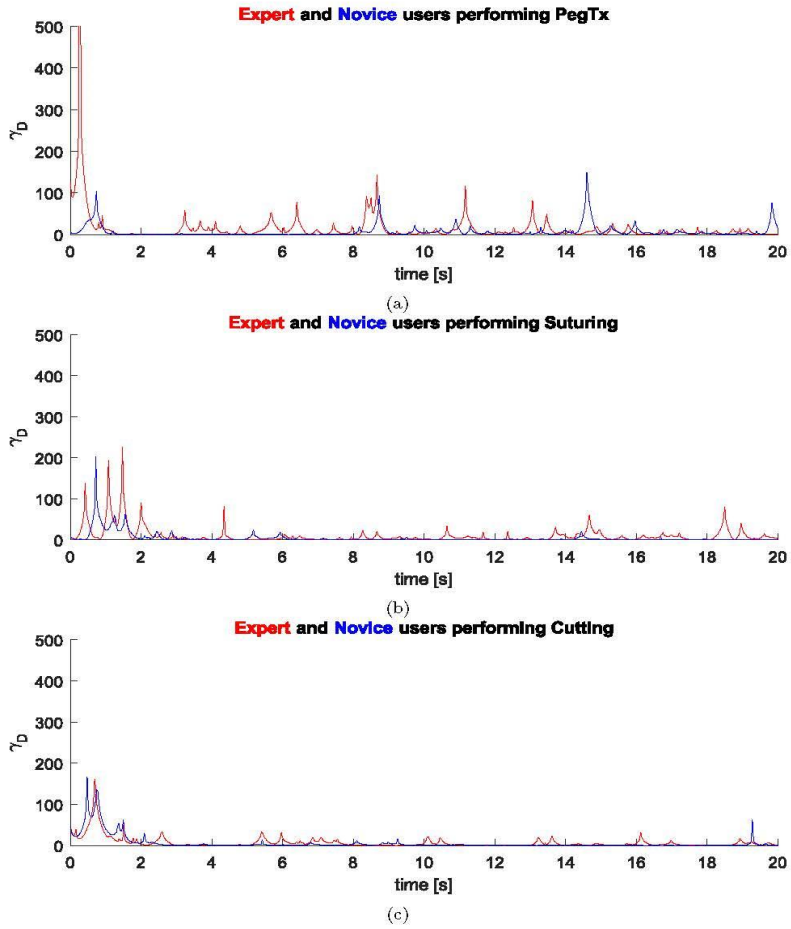
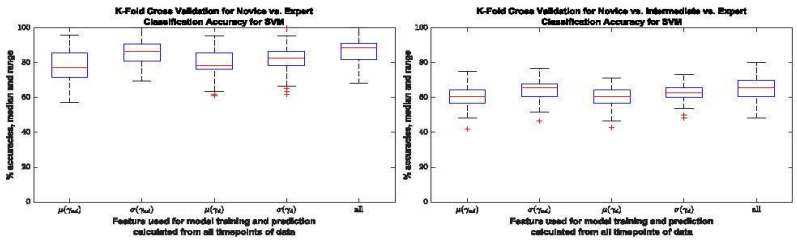


Fig. 2: Example of $\sigma(\gamma_d)$ activity over the first 20 seconds of each trial

3.2 Three Class Classification: FLS Expert, FLS Intermediate, FLS Novice

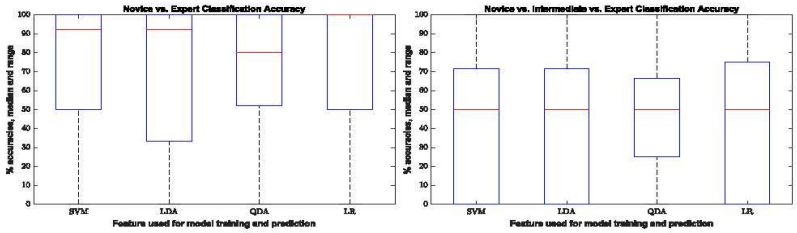
Figure 3b uses the same k-fold method as Figure 3a, however it predicts over all three classes (novice, intermediate, expert) rather than just between novice and expert. It shows a median agreement with desired skill class of 62% when trained

Seattle Children's — Internal



(a) 2-class classification (b) 3-class classification

Fig. 3: K-fold leave-trial-out SVM classification task-blind training and task-blind testing, comparing strength of each feature.



(a) 2-class classification (b) 3-class classification

Fig. 4: LOUO classification task-blind training and task-blind testing using ${}^{t}\mu(\gamma_d)$, ${}^{t}\sigma(\gamma_d)$, ${}^{t}\mu(\gamma_{nd})$, and ${}^{t}\sigma(\gamma_{nd})$ features combined to train each model, comparing strength of machinery.

using all features. Figure 4b uses the same LOUO method as Figure 4a, and it also predicts over all three classes. It shows a median agreement of 50%.

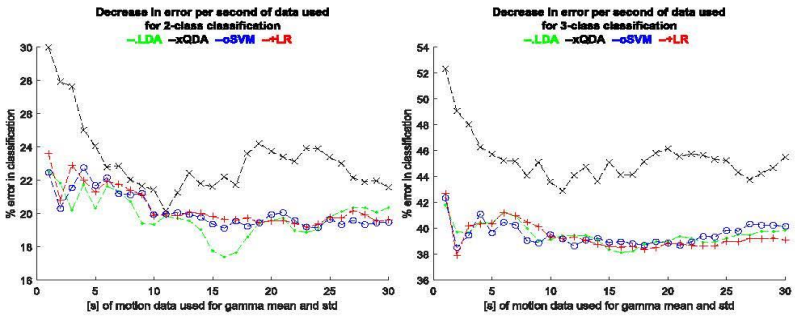
3.3 Minimum Time to Classification

Figure 5a was generated from SVM, QDA, LR and LDA models trained using all four features and shows the overall model error rate of as t is increased from 0s to 30s, which increases the number of data points in a trial used to calculate ${}^{t}\mu(\gamma_d)$, ${}^{t}\sigma(\gamma_d)$, ${}^{t}\mu(\gamma_{nd})$, and ${}^{t}\sigma(\gamma_{nd})$. Integer values of t from 1 to 30 were used. Figure 5b uses similar methods, but is classifying between novice, intermediate and expert surgeons.

3.4 Validation with Similar Methods

Figure 6a and 6b compare the performance for 2-class classification of the γ parameters against previously validated aggregate task metrics as described in [3].

Seattle Children's — Internal



(a) 2-class classification (b) 3-class classification

Fig. 5: Error rate per number of seconds used to calculate $\mu({}^t\gamma_d)$, $\sigma({}^t\gamma_d)$, $\mu({}^t\gamma_{nd})$, and $\sigma({}^t\gamma_{nd})$.

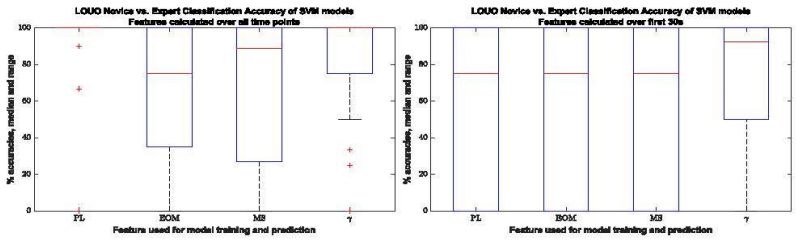
Table 2: 2-class classification accuracy results per algorithm and per task. Models generated for this figure were trained task-blind, results at left reports how well the model classifies each task

	Task Specific accuracy (μ) over all kN model accuracies in cross validation, mean(std dev)			Task Blind mean accuracy over all kN model accuracies in cross validation	
	PegTx $\mu(\sigma)$	Cutting $\mu(\sigma)$	Suturing $\mu(\sigma)$	Best % Accuracy Mean/Median	Min time to 90% of best accuracy
LDA	80.1 (.13)	86.1 (.11)	83.0 (.2)	83.7/85.7	2s
QDA	87.6 (.12)	86.9 (.11)	66.0 (.24)	82.5/82.6	4s
SVM	85.2 (.12)	90.2 (.10)	82.6 (.20)	87.2/88.7	8s
LR	84.6 (.12)	90.2 (.10)	81.5 (.21)	86.6/86.4	7s

Table 3: 3-class classification accuracy results per algorithm

	Task Specific accuracy (μ) over all kN model accuracies in cross validation, mean(std dev)			Task Blind mean accuracy over all kN model accuracies in cross validation	
	PegTx $\mu(\sigma)$	Cutting $\mu(\sigma)$	Suturing $\mu(\sigma)$	Best % Accuracy Mean/Median	Min time to 90% of best accuracy
LDA	57.3 (.13)	63.0 (.14)	66.0 (.16)	61.6/61.7	2s
QDA	66.6 (.12)	59.2 (.14)	44.3 (.16)	58.9/58.6	4s
SVM	62.2 (.13)	67.0 (.13)	65.5 (.14)	65.1/65.5	3s
LR	60.8 (.12)	68.3 (.13)	65.7 (.15)	65.1/64.3	3s

Seattle Children's — Internal



(a) Features calculated at end of task time (b) Features calculaed at 30s point in task time

Fig. 6: LOUO validation for a 2-class classification using SVM for task-blind training and task-blind testing, comparing strength of each feature set.

PL_d and PL_{nd} are the path length variables calculated for the dominant and non-dominant hand. The PL boxes represents the accuracy of an SVM trained using PL_d and PL_{nd} together as features, with accuracy measured using k-fold for N iterations. The same applies for EOM and MS boxes. The γ box represents the accuracy of an SVM trained using $\mu(\gamma_d)$, $\sigma(\gamma_d)$, $\mu(\gamma_{nd})$, and $\sigma(\gamma_{nd})$.

4 Conclusion

The results give support to our three initial hypotheses. The first hypothesis is supported by Figures 1a - 1d and Table 1. The feature $\sigma(\gamma_d)$ is taken to represent the intra-subject variability in the γ_d parameter for a hand. A low intra-subject γ_d (i.e. a small $\sigma(\gamma_d)$), may imply a given subject is nearly following the motion law outlined in [9], where the tool tip is assumed to maintain a constant γ_d. Broadening the scope to how these skill levels behave at the group level, from Table 1 it is observable that novice subjects have inter-subject median and standard deviation values for $\sigma(\gamma_d)$ and $\sigma(\gamma_{nd})$ that are comparatively lower than experts. This supports the first hypothesis. A comparatively low inter-subject standard deviation for the $\sigma(\gamma_d)$ feature for the novices may imply a behavioral pattern between subjects. Meanwhile, a comparatively low inter-subject median for the $\sigma(\gamma_d)$ feature for the novices may imply that, as a group, novices stay closer to the "natural" motion law. This could suggest that experienced surgeons mature out of this adherence to the motion pattern with practice for laparoscopic tools.

Second, the scalar parameters $\sigma(\gamma_d)$, $\mu(\gamma_d)$,$\sigma(\gamma_n d)$, and $\mu(\gamma_n d)$ were able to train LDA, QDA, SVM and LR models to predict the class of partitioned data with mean cross validation accuracy in the 85% region for binary classification and in the 60% region for ternary classification. Prediction accuracy using LOUO yielded median accuracy of up to 100% for binary classification and 50% median accuracy for ternary classification using Logistic Regression. It should be noted that there is a large variation in classification accuracy across the different users for the LOUO box plots. Chiefly, for a 2-class classification the 25th percentile is as low as 33% for certain users while a 3-class classification gets all the way to 0% for certain users. The outliers also reach 0% for 2-class classification. Further

Seattle Children's — Internal

investigation will have to be made into this behavior. It is unclear at this point whether this exposes a limitation in the chosen features or whether our data still has insufficient N to capture human variability in surgery.

Third, Figure 5a, Figure 5b, Table 2 and Table 3 show that for all four tested algorithms, the time required to get within 90% of the best observed accuracy is less than the full task duration. Note there is some oscillation in several of the curves in Figure 5a. It is not obvious what this signifies, e.g. data may be truncated through incomplete maneuvers or γ may be only significant at sustained speeds to rise above the noise floor. Exploring this will require a dataset with motion segments continuously labeled by skill level.

In addition, this method was compared against previously validated methods. For the EDGE dataset, Figures 6a and 6b show that γ parameters outperform economy of motion and motion smoothness, but fall short of path length. When measured short of task completion (at 30s), Figure 6b shows that γ parameters outperform the validation features. This is expected, since the validation metrics are heavily influenced by task time, a characteristic γ features are free of.

Determining the FLS class in the first seconds of a task for this dry-lab simulation data is a significant outcome. This implies that a trainee and proctor can potentially take less time for FLS certification. Also, traditional human-required tasks in FLS penalty scoring that were resource intensive – such as counting object drops or measuring cut accuracies – may not always be required. The positive classification results that were done in task-blind settings also suggest that the isogony measure may be capturing some of the aspects of skill evident in human motion that may be obvious to expert reviewer but difficult to articulate– aspects that may allow them to infer skill from only a few seconds of a video.

Prior art has typically not investigated task-blind skill classification methods. Our results of median 85% accuracy for novice-expert classification within the first 30s of a task rival or outperform existing, often more complex approaches. The fact that a task-blind model can be generated using γ with the demonstrated accuracy suggests that γ provides insight into some task invariant attributes of skill level.

There were several limitations in this study that should be addressed in future work. This study used only dry-lab laparoscopic simulation tasks, which do not necessarily mimic real surgical maneuvers. This limits our results and conclusions to only this simulated manual laproscopy context. The skill groups used here are defined based on FLS score only and are thresholded based on the subjects available in the dataset. Data defining skill based on the surgical panel and crowd sourced skill determinations will be used in its place in the future. It was assumed that meaningful motion was occurring in the data used this study, and selecting only the first 30s of task execution was used as a surrogate for capturing meaningful motion. It is possible that some trials may include subjects keeping their tools immobile while planning their maneuvers at the start of the task, so this must be filtered out in future work. Additional datasets could also be generated by sub-sampling randomized time intervals from existing tests. This would also help investigate the question of whether the quicker diagnoses (within 2 seconds) are due to the fact that expert surgeons get to work more quickly and confidently early in the task than novices.

We do not claim that our approach, as given, is immediately useful to surgical trainees. However, it is a necessary step towards achieving formative feedback. Namely, if a skill measuring feature only correlates with task time (e.g. FLS score

Seattle Children's — Internal

is almost identical to task time [7]), it would have little or no value for formative feedback (or even as a summary metric itself). We show that isogony provides some accuracy in measuring skill even within the first N seconds, this suggests that it has some utility over task time. However, this is a necessary but not sufficient step for formative feedback. For example, a mapping of isogony features to easy-to-understand continuous motion quality scores on, say, a percentage scale could be more useful.

We conclude that predicting final FLS score from roughly the first 10 seconds of a trial is potentially feasible and that isogony provides some useful task-blind skill-classification information above simple task-time or FLS score.

Compliance With Ethical Standards

Conflicts of Interest The authors declare that they have no conflict of interest.
Ethical Standard All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.
Funding This work was supported, in part, by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-15-2-0030. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the Department of Defense.
Informed Consent Informed consent was obtained from all individual participants included in the EDGE study.

References

1. Ahmidi N, Poddar P, Jones JD, Vedula SS, Ishii L, Hager GD, 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
2. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ (2013) Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine 369(15):1434–1442
3. Chmarra M, Kolkman W, Jansen F, Grimbergen C, Dankelman J (2007) The influence of experience and camera holding on laparoscopic instrument movements measured with the trendo tracking system. Surgical endoscopy 21(11):2069–2075
4. Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, Meakins JL (1998) Development of a Model for Training and Evaluation of Laparoscopic Skills. The American Journal of Surgery 175(6):482–487
5. Fried GM (2008) FLS Assessment of Competency Using Simulated Laparoscopic Tasks. Journal of Gastrointestinal Surgery 12(2):210–212, DOI 10.1007/s11605-007-0355-0, URL http://dx.doi.org/10.1007/s11605-007-0355-0
6. Kowalewski TM (2012) Real-time quantitative assessment of surgical skill. PhD thesis, University of Washington

Seattle Children's — Internal

7. Kowalewski TM, White LW, Lendvay TS, Jiang IS, Sweet R, Wright A, Hannaford B, Sinanan MN (2014) Beyond task time: automated measurement augments fundamentals of laparoscopic skills methodology. Journal of Surgical Research 192(2):329–338
8. Kowalewski TM, Comstock B, Sweet R, Schaffhausen C, Menhadji A, Averch T, Box G, Brand T, Ferrandino M, Kaouk J, Knudsen B, Landman J, Lee B, Schwartz BF, McDougall E, Lendvay TS (2016) Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills tasks. The Journal of urology 195(6):1859–1865
9. 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
10. Lin HC, Shafran I, Yuh D, Hager GD (2006) Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. Computer Aided Surgery 11(5):220–230
11. Peters JH, Fried GM, Swanstrom LL, Soper NJ, Sillin LF, Schirmer B, Hoffman K (2004) Development and Validation of a Comprehensive Program of Education and Assessment of the Basic Fundamentals of Laparoscopic Surgery. Surgery‘ 135:21–27
12. Rogers GM, Oetting TA, Lee AG, Grignon C, Greenlee E, Johnson AT, Beaver HA, Carter K (2009) Impact of a structured surgical curriculum on ophthalmic resident cataract surgery complication rates. Journal of Cataract & Refractive Surgery 35(11):1956–1960
13. Shute VJ (2008) Focus on formative feedback. Review of educational research 78(1):153–189
14. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM (2010) Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room – a randomized controlled trial. The American journal of surgery 199(1):115–120

Seattle Children's — Internal

Appendix B

Int J CARS
DOI 10.1007/s11548-017-1610-9

ORIGINAL ARTICLE

The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data

Rodney L. Dockter[1] · Thomas S. Lendvay[2] · Robert M. Sweet[3] · Timothy M. Kowalewski[1]

Received: 27 January 2017 / Accepted: 8 May 2017
© CARS 2017

Abstract
Purpose Minimally invasive surgery requires objective methods for skill evaluation and training. This work presents the minimally acceptable classification (MAC) criterion for computational surgery: Given an obvious novice and an obvious expert, a surgical skill evaluation classifier must yield 100% accuracy. We propose that a rigorous motion analysis algorithm must meet this minimal benchmark in order to justify its cost and use.
Methods We use this benchmark to investigate two concepts: First, how separable is raw, multidimensional dry laboratory laparoscopic motion data between obvious novices and obvious experts? We utilized information theoretic techniques to analytically address this. Second, we examined the use of intent vectors to classify surgical skill using three FLS tasks.
Results We found that raw motion data alone are not sufficient to classify skill level; however, the intent vector approach is successful in classifying surgical skill level for certain tasks according to the MAC criterion. For a pattern cutting task, this approach yields 100% accuracy in leave-one-user-out cross-validation.
Conclusion Compared to prior art, the intent vector approach provides a generalized method to assess laparoscopic surgical skill using basic motion segments and passes the MAC criterion for some but not all FLS tasks.

✉ Rodney L. Dockter
dockt036@umn.edu

1 Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN, USA

2 Department of Urology, Seattle Children's Hospital, Seattle, WA, USA

3 Department of Urology, University of Washington, Seattle, WA, USA

Published online: 18 May 2017

Keywords Surgical skill evaluation · Surgical training · Surgical motion · Laparoscopic surgery

Introduction

The fundamentals of laparoscopic surgery (FLS) were developed to evaluate and credential laparoscopic surgeons. The FLS scoring criteria are based primarily on task time and number of task errors as determined by a qualified proctor. While FLS has been shown to discriminate between expert and novice subjects [18], these measures have the potential to miss key information and overemphasize task time [13]. The challenges related to laparoscopic surgery motivate the development of objective, automated, and accurate surgical skill evaluation techniques.

Prior work on surgical skill evaluation has been widespread. One approach has utilized aggregate task measures such as task time and path length [5,6]. In [16], task level metrics were used to estimate pairwise maneuver preferences with 80% accuracy. In [9], robotic arm vibrations and interaction forces were used within a composite skill rating; however, statistical analysis showed that completion time provided the primary contribution. Another method has been to decompose surgical tasks into specific gestures or 'surgemes' [15]. Using these surgemes, models for skill can be trained using a variety of machine learning approaches. Hidden Markov models (HMMs) have been used extensively to model surgical skill level. An HMM model for various surgemes was used to classify a sequence as a particular skill level [17]. This resulted in 100% classification accuracy for leave-one-super-trial-out (LOSO) cross-validation but required manually segmented surgemes and did not report leave-one-user-out (LOUO) validation results. The results of [19] had high classification rates for LOSO cross-

Seattle Children's — Internal

validation, but these results fell precipitously under LOUO validation suggesting overfitting. Another method utilizes descriptive curve coding (DCC) in which the principal direction changes within a trajectory are encoded as a string of integers [1]. With this approach, encoded common strings or motifs were used to model skill level. This method results in 98% accuracy for LOSO validation but around 90% for LOUO. Task-specific motion models have been proposed for procedures such as septoplasty [2]. This approach involves stroke-based features designed to assess the consistency and efficiency with which a surgeon removes skin from underlying cartilage. This approach gave a LOSO classification accuracy of 90%, but its applicability to other procedures is not yet clear. The ribbon area measure treats the surgical tool wrist as a brush and measures the accumulated surface area of the trajectory as a surrogate for dexterity [11]. This approach resulted in an 80% binary classification accuracy. Both the stroke-based features and ribbon area approaches are conceptually similar to the work presented here; however, we attempt to use these concepts in a manner more generalizable across tasks and that results in a higher classification accuracy. The gap in prior art has been a fully automated algorithm which provides 100% classification between obvious expert and novice surgeons using LOUO cross-validation.

Prior art has revealed a secondary problem: Data set categories are unreliably labeled relative to true skill level. These categories are typically defined by subject demographics such as caseload, academic rank, or experience level. Yet even an expert surgeon can exhibit skill decay and demonstrate a variance in skill level within a given context. True experts or technical masters can sometimes (e.g., for a given grasp or motion within an entire procedure) exhibit novice-like motions. Kowalewski et al. [14] showed that expert categories based on these demographics are unsuitable for validation studies as they often result in recorded trials from perceived experts that can exhibit poor technical skill. Overall this can confound supervised classifiers that assume a clean ground truth for correct analysis. The current gold standard for skill assessment is blinded review of surgical videos by panels of expert surgeons using structured survey tools such as the objective structured assessment of technical skill (OSATS) [7]. Birkmeyer et al. [3] showed that using similar evaluation methods technical skill can be linked directly with patient outcomes. To this end, Kowalewski et al. [13] defined a ground truth expert trial (a single recording by a given individual) as one that is deemed an expert by a consensus of three validated methods: demographically-derived expertise, FLS score, and OSATS-like video review.

We herein introduce the minimally acceptable classification (MAC) criterion for computational skill evaluation: Given an obvious novice and an obvious expert, the classification accuracy must be 100%. Some misclassification may be acceptable between other skill levels, e.g., experts versus Master or Intermediate versus expert, but not an *obvious* novice versus *obvious* expert. Here we define obvious novices as subjects who should never be allowed to operate (always disqualified) and obvious experts as subjects who should never be disqualified from operating. Surgery requires this stipulation given that patently unqualified surgeons endanger lives. Often, such a large difference is very evident via task time or a casual viewer watching a video [4]. Therefore, a rigorous motion analysis algorithm should meet this minimal performance benchmark in order to justify cost and use. While this is not a sufficient criteria, it does provide a *minimal* necessary criterion to use as a baseline in this field. Our approach in this study was twofold. First, we asked 'how valuable is raw tool motion data alone in classifying skill given the MAC criterion?' Second, we present the 'intent vectors' feature and classification scheme applied to laparoscopic tool motion. We tested the hypothesis that intent vectors successfully classify skill according to the MAC for specific tasks.

Methods

In this section, we present the data set utilized in this study, the separability analysis used to assess raw surgical motion data, and the intent vectors derivation. The lack of separability in the raw data motivates the intent vectors.

Data set

This study utilized a previously recorded data set [13] where the electronic data generation for evaluation (EDGE) platform (Simulab Corp., Seattle, WA, USA) was used to collect task video data and tool motion data from participants including surgical faculty, residents, and fellows. Participants in the study performed a subset of the FLS tasks; peg transfer, pattern cutting, and intracorporeal suturing. Each subject was asked to complete, at minimum, three iterations of the peg transfer task, two iterations of the pattern cutting task, and two iterations of the suturing task. The subject pool consisted of 98 total subjects from a variety of specialties including General Surgery, Urology, and Gynecology spanning three teaching hospitals. Two FLS-certified graders manually recorded task errors, and task completion time was automatically recorded. Task errors and completion time were then used to compute an overall FLS score for each iteration.

From this data set, we have chosen the ground truth expert group (determined by a combination of caseload, FLS score, and p-OSATS score) for our 'obvious expert' category and the FLS novice group (determined by the bottom 15th percentile of FLS scores for trials in each task) for our 'obvious novice' category. Individuals with such low scores would fail FLS and thus not be allowed to operate. The complete data

Seattle Children's — Internal

Table 1 FLS trials by task and skill level

Skill level	Peg transfer	Pattern cutting	Suturing
'Obvious novice'	29	25	13
'Obvious expert'	6	10	8

set contains 447 recorded trials across three tasks [13]. We selected only 91 of the original recorded trials to represent the extremes of 'obvious experts' and 'obvious novices.' Each trial was performed by a different subject (Table 1).

Each task was recorded with time synchronized video and tool motion data. This provided time-stamped Cartesian positions (x, y, z in cm) along with tool roll and grasper jaw angle (θ, degrees) at 30 Hz. This allowed subsequent computation of motion derivatives such as velocity and acceleration. In post-processing, surgical tool motion was segmented into distinct motions within each task based on information from the tool grasper at the distal end. A segment was considered to begin when the grasper was opened ($\theta > 3°$) and the force within the grasper jaws fells below a threshold ($F_g < 4N$). The segment was then considered complete when the jaws were closed ($\theta < 3°$) and the force applied within the grasper jaws rose above a threshold ($F_g > 4N$) for 200 ms [13]. Each tool is segmented separately, allowing for overlapping segments between each instrument (hand). The mean number of segments per trial and the mean segment duration are given in Table 2.

Functionally this segmentation scheme results in segments where a tool is moved in a trajectory toward an object, and then the jaws are closed around the object to secure it, thus ending the segment. Our segments focused only on tool motion where the surgeon is reaching toward an object (e.g., before grasping or cutting), a motion which is prevalent in nearly all surgical tasks. The goal of this segmentation scheme was to be generalizable to all surgical tasks as compared to task-specific surgical gestures. We expected that some spurious false positives may occur within segmentation and assumed that these false segments occur equally across skill groups.

Value of 'raw motion data' for classification

To explore the separability of dexterous skill levels given raw motion data from EDGE, we refined and utilized information theoretic techniques, starting with the RELIEFF algorithm [12]. This is used in binary classification to rank features based on their ability to separate the data effectively. For each point, we find the K-nearest neighbors belonging to the true class (hit) and the opposite class (miss). Using these nearest neighbors, a mean distance to both the hit neighbors (D_{hit}) and the miss neighbors (D_{miss}) is computed. The weights for a particular feature (W_f) are updated according to the difference between mean hit distance and mean miss distance (computed using that particular features data) (Eq. 1).

$$W_f = \sum_{i=1}^{N} \left(D_{\text{hit}_i} - D_{\text{miss}_i} \right). \tag{1}$$

Once weights for each feature have been computed, the features are sorted based on weight. Features with the highest weights are considered the most relevant features for classification. RELIEFF and its variants are limited to considering each feature separately and do not consider combinations of features simultaneously.

An obvious extension of the RELIEFF approach for multiple features (a variant termed RELIEF-RBF) utilizes radial basis functions (RBF) to estimate the probability density function given within class (hit) and between class (miss) data across any combination of n dimensions. As compared to the standard RELIEFF approach, all data from all dimensions contribute to the overall probability of that data point instead of only considering nearby neighbors in a single dimension. A training data set is utilized, and each point (indexed by i) within the n-dimensional set is assigned a probability estimate via RBFs for within class probability (P_{hit}) and between class probability (P_{miss}) (Eqs. 2, 3).

$$P_{i,\text{hit}} = \frac{\sum_{j=1}^{N_{\text{hit}}} e^{-(\epsilon \|x_i - x_j\|)^2}}{N_{\text{hit}}} \tag{2}$$

$$P_{i,\text{miss}} = \frac{\sum_{k=1}^{N_{\text{miss}}} e^{-(\epsilon \|x_i - x_k\|)^2}}{N_{\text{miss}}}. \tag{3}$$

The bandwidth variable ϵ is used to scale the kernel radius given a standard deviation. Given the class-specific probability estimates for each data point, we compute the relative separability of each data point between its hit class and miss class. This requires computing the Kullback–Leibler (KL) divergence of each point using both probability estimates (Eq. 4).

$$W_{i,\text{rbf}} = P_{i,\text{hit}} \cdot \log\left(\frac{P_{i,\text{hit}}}{P_{i,\text{miss}}} \right). \tag{4}$$

Table 2 Mean segment count ± standard deviation and [mean segment duration] by task and skill level

Skill level	Peg transfer	Pattern cutting	Suturing
'Obvious novice'	30.5 ± 4.6 [260 ms]	61.9 ± 18.1 [130 ms]	41.7 ± 18.3 [203 ms]
'Obvious expert'	24.8 ± 1.3 [105 ms]	27.1 ± 4.5 [68 ms]	12.4 ± 3.0 [107 ms]

Seattle Children's — Internal

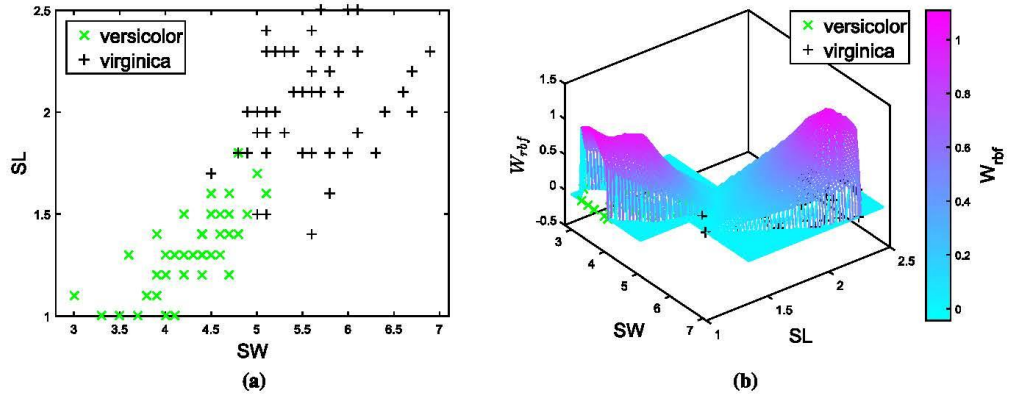


Fig. 1 RELIEF-RBF for sepal width (SW) and sepal length (SL) data from the Versicolor and Virginica classes. **a** 2D Fisher Iris data. **b** RELIEF-RBF weights

Each data point x_i in d-dimensional space ($d \leq n$) is assigned an estimate of separability $W_{i,\text{rbf}}$ (i.e., relevance in terms of classification use). The mean relevance weighting from all points in the training data set yields an aggregate estimate of the relevance weighting for that combination of features. This relevance weight is then compared with other combinations to improve feature selection for large, multidimensional, numerical data sets. A two-dimensional example of the relevance weights for two classes of the Fisher Iris data set [8] (Versicolor and Virginica) is given in Fig. 1. The RELIEF-RBF algorithm rewards only regions with high confidence of separability (high $W_{i,\text{rbf}}$), while penalizing *both* regions with a prevalence of all classes *and* regions that are data scarce (low $W_{i,\text{rbf}}$).

In both RELIEFF and RELIEF-RBF, all dimensions are mean–variance pre-scaled to account for data range effects. The weights for both methods are un-normalized and are used to compare the relative separability across dimensions.

Using both RELIEFF and the RELIEF-RBF, we investigated which states from the raw EDGE motion data had the highest separability. The states used in this study are given in Eq. (5) where $\dot{x}, \dot{y}, \dot{z}$ terms represent derivatives w.r.t. time of the Cartesian location of the surgical tool tip. χ_t is sample at each time step in the data set. The Cartesian position of the surgical tool $[x, y, z]$ was excluded because of its relationship to the present surgical gesture. All resulting feature combinations were investigated.

$$\chi_t = \left[\theta \dot{\theta}\ \dot{x}\ \dot{y}\ \dot{z}\ \ddot{x}\ \ddot{y}\ \ddot{z}\ \dddot{x}\ \dddot{y}\ \dddot{z}\ \|\dot{x},\ \dot{y},\ \dot{z}\|\|\ddot{x},\ \ddot{y},\ \ddot{z}\|\right]. \tag{5}$$

For comparison, we also applied RELIEF-RBF to the Fisher Iris data set, a well known, separable data set. Using the three surgical motion states with the highest RELIEF-RBF separability, we employed a random forest classification (100 trees) to examine the classification accuracy in a LOUO cross-validation scheme.

Intent vectors

We present a novel motion statistic for surgical skill classification. The 'intent vectors' statistic is based on the overall goal of a motion segment. Using the starting and ending location of a motion segment as endpoints, we compute a vector which represents the ultimate goal of that segment. We assume this intent vector is the ideal line of motion for a given segment; then we compute metrics which represent the amount of deviation from this optimal trajectory.

For a segment of Cartesian tool position data of length N, we have $\Psi = [D_1, D_2, \ldots, D_N]$ where $D_i = [x, y, z]$ represents the 3D location at time $t = i$. The intent vector is then computed in Eq. (6).

$$\overrightarrow{\text{IV}} = \frac{D_N - D_1}{\|D_N - D_1\|}. \tag{6}$$

From this intent vector, the progress of each point in Ψ along this line can contextualize other actions relative to the ultimate trajectory. The intent vector progress value (IVP) is computed according to Eq. (7) using a dot product operator and scaled by the magnitude of the intent vector (thus fixing the starting and ending points at 0 and 1). An illustrative example is given in Fig. 2a.

$$\text{IVP}_i = \frac{(D_i - D_1) \cdot \overrightarrow{\text{IV}}}{\|D_N - D_1\|}. \tag{7}$$

From the intent vector framework, we also compute the intent vector angle (IVA): the angle of motion relative to the

Seattle Children's — Internal

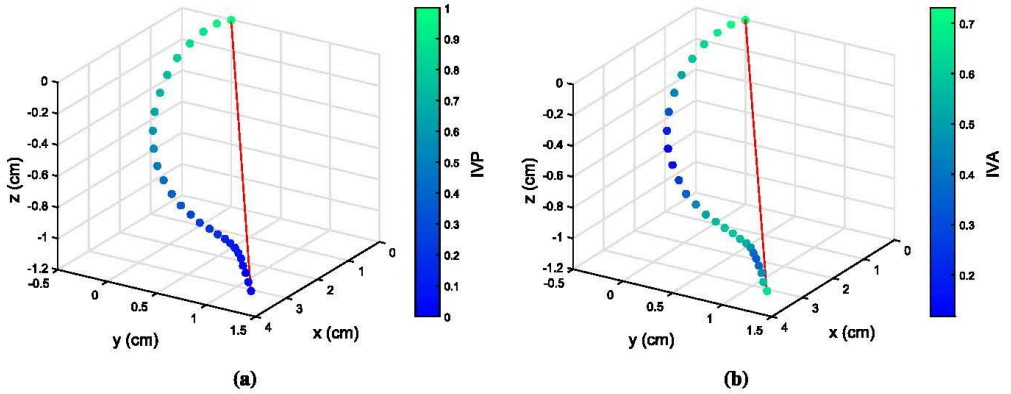


Fig. 2 Intent vector measures. **a** Intent vector progress in 3D. **b** Intent vector angle in 3D

overall angle of the intent vector. IVA is computed for each point in Ψ by taking the difference at a given point in time between the current tool location and the previous location $(D_i - D_{i-1})$ which is then normalized to give a unit vector in 3D space (S_i). Given this instantaneous unit vector, we compare with the overall intention, indicating the degree to which the tool is moving in the correct direction or doubling back (Eqs. 8, 9).

$$S_i = \frac{D_i - D_{i-1}}{\|D_i - D_{i-1}\|} \tag{8}$$

$$\text{IVA}_i = \cos^{-1}(S_i \cdot \vec{\text{IV}}). \tag{9}$$

The value of IVA is bounded between $0 < \text{IVA} < \pi$ since we are not concerned with the direction that the angle differs from the overall intent. An illustrative example is given in Fig. 2b. The intent vector framework was implemented for all motion segments within the EDGE data set. For each task, the IVA and IVP measures were compiled into a 2D feature vector with corresponding skill labels. A plot of IVA and IVP for the suturing task can be found in Fig. 4a.

Given the high-degree of similarity in the intent vector space, to use the intent vector data within a classification scheme we employed a classification approach which focuses on deviations from the region of high expert probability. We first identified the region in 2D IVA–IVP space with the highest density of expert surgical motion. We employed a modified version of the RELIEF-RBF algorithm and threshold the relevance weights for the expert class (Eq. 10).

$$W_{i,\text{exp}} = P_{i,\text{exp}} \cdot \log\left(\frac{P_{i,\text{exp}}}{P_{i,\text{nov}}}\right). \tag{10}$$

Here $W_{\text{exp}} = W_{\text{rbf}}$ from Eq. (4) where expert is the hit class. All training data are assigned a relevance weight relative to the expert data. A threshold on $W_{i,\text{exp}}$ is computed using an information gain maximization similar to the typical decision stump algorithm [10]. We identify a threshold (T_w) such that classification of the intent vector data follows Eq. (11) and maximizes the information gain $(\text{IG} = H(Y|X) - H(Y))$ for classification $(Y = \text{skilllevel})$ given $(X = [\text{IVA}, \text{IVP}])$.

$$Y = \begin{cases} \text{Novice}, & W_{\text{exp}}(X) < T_w \\ \text{Expert}, & W_{\text{exp}}(X) \geq T_w. \end{cases} \tag{11}$$

Using the relevance weight threshold, we retain all expert data in [IVA, IVP] space above T_w as 'true expert data' and train a Gaussian probability model for online classification $(P_{\text{exp}}(X|\mu, \sigma))$. A threshold value for this Gaussian model (T_p) is found by taking the $P_{\text{exp}}(X)$ at the minimum $W_{i,\text{exp}}(X) > T_w$ value.

The next step is to classify each individual time-indexed data point within a given segment for a specific surgeon. For surgeon (g) and segment (s), the time series data are given as $\Lambda_{g,s} = [\lambda_1, \lambda_2, \ldots, \lambda_N]$ where $\lambda_i = [\text{IVA}, \text{IVP}]$ at time $t = i$. Using $P_{\text{exp}}(X|\mu, \sigma)$, we classify each data point as 1 or 0 to signify novice or expert, respectively (Eq. 12). Values where $y_i = 1$ are considered a 'demerit' for behaving like a novice and are used in the overall evaluation of the motion.

$$y_i = \begin{cases} 1, & P_{\text{exp}}(\lambda_i) < T_p \\ 0, & P_{\text{exp}}(\lambda_i) \geq T_p. \end{cases} \tag{12}$$

Given a vector of time-indexed motion demerits $q_{g,s} = [y_1, y_2, \ldots, y_N]$, we compute a mean score for that particular segment $\text{SK}_{g,s} = \text{mean}(q_{g,s})$. Given the 1, 0 labels, this score has the effect of being very low for frequent expert motions and higher if motions fall outside the 'true expert'

Seattle Children's — Internal

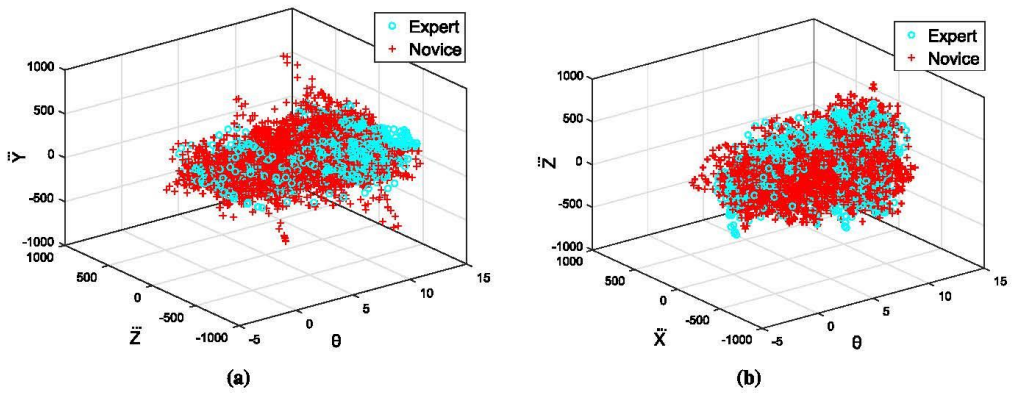


Fig. 3 Relevance weightings for raw motion states. **a** Top three RELIEFF states. **b** Top three RELIEF-RBF states

model (many novice demerits). We train a threshold based on the average SK scores (T_{sk}) for expert and novice surgeons using a decision stump approach. We employ a LOUO scheme per skill group (LOUOpG) (i.e., leave one obvious novice and one obvious expert out per training) and test each left-out surgeon based on all motion segments (Eq. 13).

$$C_g = \begin{cases} \text{Novice}, & \text{mean}(\text{SK}_{g,s}) > T_{sk} \\ \text{Expert}, & \text{mean}(\text{SK}_{g,s}) \leq T_{sk}. \end{cases} \tag{13}$$

For each LOUOpG iteration, we recompute all relevant measures and thresholds, i.e., W_{exp}, T_w, T_p, and T_{sk} based on the training data set alone, therefore limiting overfitting for the validation data.

In order to compare the accuracy of our classification approach, we utilized previously validated aggregate task metrics as highlighted in [5]. For this comparison, we used a feature vector comprised of tool path length, economy of motion (Eq. 14), motion smoothness, and motion curvature (Eq. 15, where $\dot{r} = \|\dot{x}, \dot{y}, \dot{z}\|$) ($\bar{\chi} = [\text{PL, EOM, MS, MC}]$). A linear discriminant analysis (LDA) classifier (class-based means and covariances, equal weighting) was trained on this feature vector to classify skill levels. We again employed a LOUOpG cross-validation with this classifier. We also examined classification using a combination of intent vectors and aggregate metrics with combined feature vector $\hat{\chi} = [\bar{\chi}, \text{mean}(\text{SK}_{g,s})]$. Again we utilized a standard LDA classifier in a LOUOpG cross-validation to classify a complete task.

$$\text{EOM} = \frac{\text{Path Length}}{\text{Task Time}} \tag{14}$$

$$\text{MC} = \frac{\dot{r} \times \ddot{r}}{|\dot{r}|^3}. \tag{15}$$

Results

Value of 'raw motion data' for classification

The relevance of the raw motion states was examined for all states in Eq. (5). The three motion states with the highest relevance weights according to RELIEFF were found to be $[\theta, \dddot{z}, \dddot{y}]$. The corresponding RELIEFF weights were $[2.3 \times 10^{-3}, 2.7 \times 10^{-3}, 3.0 \times 10^{-3}]$. A plot of these three states is given in Fig. 3a.

RELIEF-RBF gave slightly different states with high relevance. The motion states with the highest relevance weights according to RELIEF-RBF were found to be $[\theta, \dddot{x}, \dddot{z}]$. The corresponding RELIEF-RBF weight was 6.7×10^{-3} for this combination of states. A plot of these three states is shown in Fig. 3b. The additional relevance weights for the other motion states are not included for the sake of brevity but were all similarly low.

All states in the motion data had separability measures that were orders of magnitude lower than the separability of the Fisher Iris data set, which has a maximum relevance weight of 0.63 for sepal width and sepal length (RELIEF-RBF). Using a random forest classifier on the top RELIEF-RBF motion states gave a classification accuracy of 70.5% and an out-of-bag error of 0.28. Given the relatively low feature weights for the raw motion data, the resulting classification accuracy did not fulfill the MAC criterion, being well below 100%.

Intent vectors

A sample plot of the intent vectors space is given in Fig. 4a. These data indicate clear differences between novices and experts. Novices spend far more time outside the 0–1 range of the IVP, meaning they often backtrack and overshoot the starting and ending points. Additionally, experts spend a lot

Seattle Children's — Internal

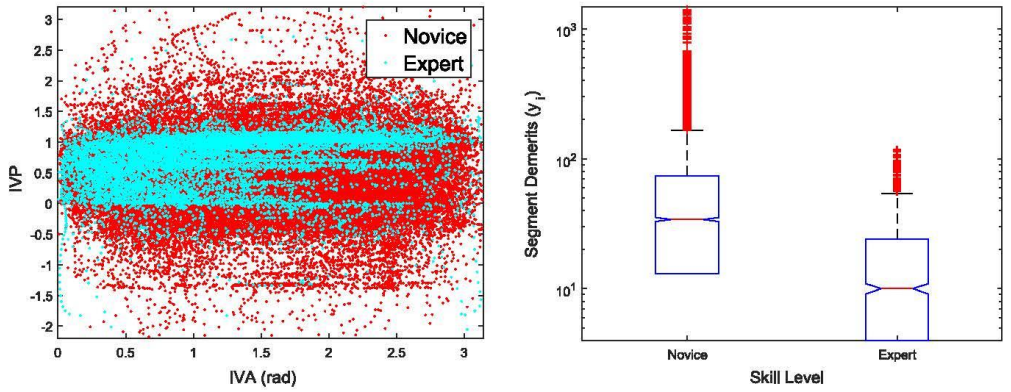


Fig. 4 Intent vector data (**a**) and demerit counts (**b**) (obvious novice and expert) for suturing task box-plot notch indicate range of 95% confidence for median separation. **a** IVA versus IVP with class labels. **b** Per-segment demerits (y_i)

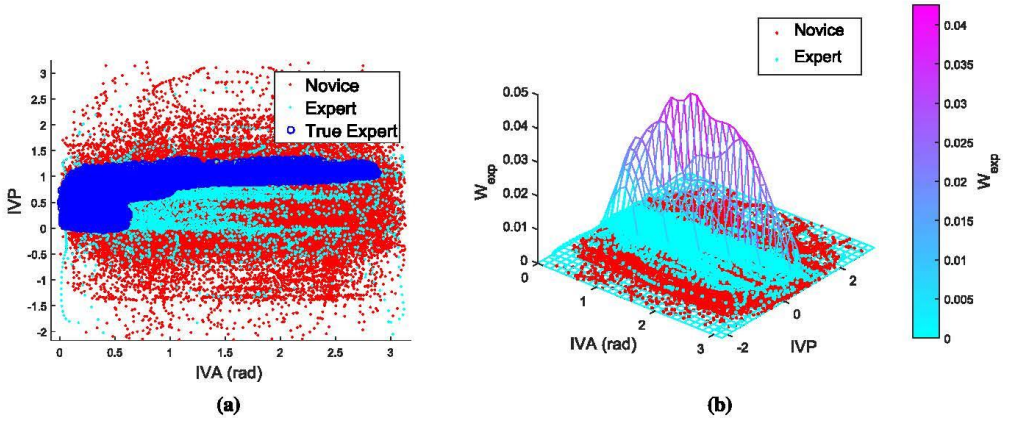


Fig. 5 Intent vector data with 'true expert' data and RELIEF-RBF weights (obvious novice and expert). **a** 'True expert' region. **b** RELIEF-RBF weights

of time with low IVA values meaning they generally head in the correct direction. However, experts also have varied IVA values around the endpoint of segments (IVP = 1), meaning that near the endpoint, experts make fine adjustments to their approach.

The intent vector classification yielded a large separation among segment demerit counts (y_i) between expert and novice surgeons. A plot of these values for each class is given in Fig. 4b. The mean segment demerit count was found to be 65.9 (std = 105.2) for novices and 22.6 (std = 27.7) for experts. The relevance weights (W_{exp}) and 'true expert' data in the intent vector space are shown in Fig. 5.

The intent vector framework yielded an average classification accuracy of 97% between novices and experts using a LOUOpG scheme for all tasks combined (Table 3). The intent vector approach fails to pass the MAC criterion for all tasks. However, it does achieve the MAC for the pattern cutting task.

An example plot of expert versus novice total segment demerits and the learned thresholds T_{sk} (Eq. 13) from all LOUOpG iterations is given in Fig. 6 for the intracorporeal suturing task. Results suggest the existence of an ideal threshold (obtainable using all available data) that provides clear separation between novice and expert data in the suturing task.

For comparison, the LDA classifier using the aggregate task metric features ($\bar{\chi}$) achieved the classification rates in square brackets in Table 3. These measures failed to achieve

Seattle Children's — Internal

Table 3 Intent vectors [aggregate metrics] {combined features} classification accuracy (%)

Skill level	Peg transfer	Pattern cutting	Intracorporeal suturing
Novice	96.5 [100[a]] {100[a]}	100[a] [96] {96}	100[a] [92.3] {92.3}
Expert	83.3 [83.3] {86.2}	100[a] [90] {100[a]}	92.3 [87.5] {100[a]}
Macro-accuracy	94.2 [97.1] {97.6}	100[a] [94] {97.2}	97.1 [90] {95.2}

[a] Achieves MAC criterion

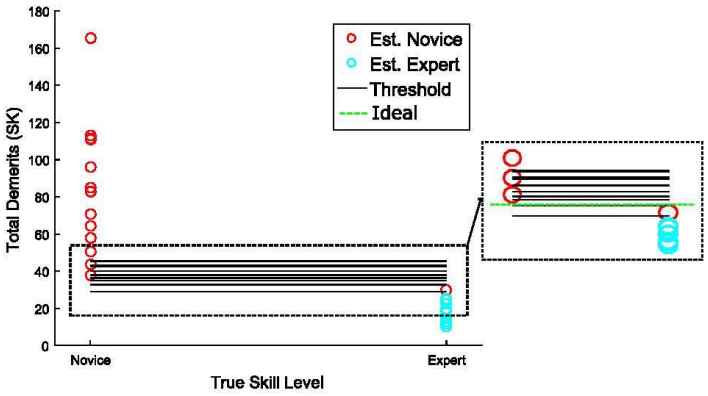


Fig. 6 LOUOpG classification using intent vectors with thresholds (T_{sk}) and ideal separable threshold

100% (macro-accuracy) classification for any of the tasks. The intent vector approach performed better than aggregate measures for both the suturing and cutting tasks, but worse in the peg transfer task. The combined feature vector $\hat{\chi}$ achieved equivalent or better macro-accuracy than the aggregate metrics alone for all tasks, indicating improved performance through the incorporation of intent vectors.

Conclusion

We presented the minimally acceptable classification (MAC) criterion for surgical skill classifiers. That is, given obvious expert and obvious novice data, a classification accuracy of 100% must be demonstrable as a minimal criteria for surgical skill classification. This requires stating both the classifier performance under LOUO-level cross-validation and enumerating its useful benefits over existing methods like summary metrics (e.g., task time).

We investigated the separability of raw tool motion data between obvious novices and experts with this MAC criteria in mind. As visible in Fig. 3, our results indicate extremely low separability—orders of magnitude lower than, say, the Fisher Iris dataset. This was true using both the RELIEFF and RELIEF-RBF feature selection algorithms. This suggests that motion data alone are statistically inseparable for classification given the MAC criterion. This is reiterated by the poor performance of the random forest classifier using raw tool motion alone. This motivates the inclusion of additional context (e.g., video data, tracking tissues, and tool–tissue interaction) to amplify the relevance of input data to the classification problem.

The intent vector feature and classifier performed surprisingly well given the observed low separability of the raw tool motion data. The overall classification rate of 97% rivals or surpasses prior the literature especially under LOUOpG cross-validation. We note that this approach fails to achieve the MAC criterion for all three FLS tasks. However, our intent vector classifier does partially succeed under the MAC criterion for two special cases: the cutting task and identifying obvious novices in the suturing task. Closer inspection in Fig. 6 reveals that the intent vector can fully separate the suturing task (and hence classify with 100% accuracy to achieve the MAC criterion) given an ideal threshold. This approach achieves equivalent or better results when compared with aggregate task metrics common in prior art. When used in the combined feature vector $\hat{\chi}$, we found that intent vectors improve classification accuracy when compared with the aggregate task metrics alone. Furthermore, for the cutting and suturing tasks, the intent vector provides additional value beyond summary metrics like task time. Notably, it returns classification results upon completion of each motion segment. This permits use cases such as (1) identifying only the worst portions of a surgical video for streamlined targeted review or (2) providing skill feedback in near real time at the completion of every motion. The segmentation approach used has the additional benefits of not requiring manual segmentation and being task agnostic.

Seattle Children's — Internal

We propose that the MAC criterion be adopted in surgical skill research as a minimal benchmark for a surgical skill classifier. Otherwise, the cost or complexity of sophisticated algorithms may not be justified. Using MAC also demands more carefully chosen ground truth skill categories to ensure accurate establishment of the ground truth, e.g., combining multiple criteria such as OSATS review, caseload, and procedural metrics. Failure to establish such a clean ground truth may hamper scientific progress in skill evaluation research.

This study has multiple limitations. This approach has only been applied to manual laparoscopic data on simulated tasks. Our conclusions may not hold for other contexts such as live surgery or robotic systems. The high selectivity of our 'obvious expert' inclusion criteria resulted in relatively small numbers of trials for cross-validation. Future work will include additional data collection to remedy this and applying the intent vector framework to ternary skill level classification. Additional analysis will investigate the concordance of intent vector metrics with FLS scores. We intend to compare our approach with the DCC and ribbon area measures [1,11]. This method has only been applied within our ballistic approach segmentation scheme; future work will investigate whether intent vectors can be applied to other actions such as needle passing. The current framework assumes the overall intent of each segment is correct and does not account for motion with incorrect intent. This segmentation scheme has the potential for false positives but is assumed to affect skill groups equally.

Acknowledgements R. Dockter was supported by the University of Minnesota Interdisciplinary Doctoral and Informatics Institute (UMII) MnDRIVE fellowships.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the EDGE study.

References

1. Ahmidi N, Gao Y, Béjar B, Vedula SS, Khudanpur S, Vidal R, Hager GD (2013) String motif-based description of tool motion for detecting skill and gestures in robotic surgery. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 26–33
2. Ahmidi N, Poddar P, Jones JD, Vedula SS, Ishii L, Hager GD, Ishii M (2015) Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. Int J Comput Assist Radiol Surg 10(6):981–991
3. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442
4. Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B, Kuksenok K, Aragon C, Holst D, Lendvay T (2014) Crowd sourced assessment of technical skills: a novel method to evaluate surgical performance. J Surg Res 187(1):65–71
5. Chmarra MK, Klein S, de Winter JC, Jansen FW, Dankelman J (2010) Objective classification of residents based on their psychomotor laparoscopic skills. Surg Endosc 24(5):1031–1039
6. Datta V, Mackay S, Mandalia M, Darzi A (2001) The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. J Am Coll Surg 193(5):479–485
7. Faulkner H, Regehr G, Martin J, Reznick R (1996) Validation of an objective structured assessment of technical skill for surgical residents. Acad Med 71(12):1363–1365
8. Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188
9. Gomez ED, Aggarwal R, McMahan W, Bark K, Kuchenbecker KJ (2016) Objective assessment of robotic surgical skill using instrument contact vibrations. Surg Endosc 30(4):1419–1431
10. Iba W, Langley P (1992) Induction of one-level decision trees. In: Proceedings of the ninth international conference on machine learning, pp 233–240
11. Jog A, Itkowitz B, Liu M, DiMaio S, Hager G, Curet M, Kumar R (2011) Towards integrating task information in skills assessment for dexterous tasks in surgery and simulation. In: 2011 IEEE international conference on robotics and automation (ICRA). IEEE, pp 5273–5278
12. Kononenko I, Šimec E, Robnik-Šikonja M (1997) Overcoming the myopia of inductive learning algorithms with RELIEFF. Appl Intell 7(1):39–55
13. Kowalewski TM, White LW, Lendvay TS, Jiang IS, Sweet R, Wright A, Hannaford B, Sinanan MN (2014) Beyond task time: automated measurement augments fundamentals of laparoscopic skills methodology. J Surg Res 192(2):329–338
14. Kowalewski TM, Sweet R, Lendvay TS, Menhadji A, Averch T, Box G, Brand T, Ferrandino M, Kaouk J, Knudsen B, Landman J, Leek B, Schwartz BF, McDougall E (2016) Validation of the AUA BLUS tasks. J Urol 195(4):998–1005
15. Lin HC, Shafran I, Yuh D, Hager GD (2006) Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions. Comput Aided Surg 11(5):220–230
16. Malpani A, Vedula SS, Chen CCG, Hager GD (2014) Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In: International conference on information processing in computer-assisted interventions. Springer, pp 138–147
17. Reiley CE, Hager GD (2009) Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In: Medical image computing and computer-assisted intervention—MICCAI 2009. Springer, pp 435–442
18. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM (2010) Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room a randomized controlled trial. Am J Surg 199(1):115–120
19. Tao L, Elhamifar E, Khudanpur S, Hager GD, Vidal R (2012) Sparse hidden markov models for surgical gesture classification and skill evaluation. In: International conference on information processing in computer-assisted interventions. Springer, pp 167–177

Seattle Children's — Internal

Appendix C

Laparoscopic Skill Classification using the Two-Third Power Law and the Isogony Principle

Anna French[1]
Timothy M. Kowalewski[1], **PhD**

[1]Department of Mechanical Engineering
University of Minnesota

1 Background

Surgical skill evaluation is a field that attempts to improve patient outcomes by accurately assessing surgeon proficiency. An important application of the information gathered from skill evaluation is providing feedback to the surgeon on their performance. The most commonly utilized methods for judging skill all depend on some type of human intervention. Expert panels are considered the gold standard for skill evaluation, but are cost prohibitive and often take weeks or months to deliver scores. The Fundamentals of Laparoscopic Surgery (FLS) is a widely adopted surgical training regime. Its scoring method is based on task time and number of task-specific errors, which currently requires a human proctor to calculate. This scoring method requires prior information on the distribution of scores among skill levels, which creates a problem any time a new training module or technique is introduced. These scores are not normally provided while training for the FLS skills test, and [1] has shown that FLS scoring does not lend any additional information over sorting skill levels based on task time. Crowd sourced methods such as those in [2] have also been used to provide feedback and have shown concordance with patient outcomes, however it still takes a few hours to generate scores after a training session.

It is desired to find an assessment method that can deliver a score immediately following a training module (or even in real time) and depends neither on human intervention nor on task-specific probability distributions. It is hypothesized that isogony-based surgical tool motion analysis discerns surgical skill level independent of task time.

2 Methods

2.1 Data Set

This study used tool motion data gathered from the EDGE (Electronic Data Generation and Evaluation) study [3]. This dataset contains 295 different samples of surgeons at varying skill levels interacting with a dry-lab surgical training environment performing 108 peg transfer (PegTx), 63 suturing and 124 circle cutting tasks.

Each sample is composed of a video recording of the training module (30 Hz), the cartesian space laparoscopic tool motion data corresponding to each video frame (30 Hz), features such as task name, and an FLS skill rating. The tool tip position and velocity measurements from the tool motion data were used to calculate our features of interest for evaluating skill. The FLS score was mapped in [1] to a ternary ranking of subjects as novice, intermediate or expert. The trials utilized for this experiment include 157 FLS novices, 71 FLS intermediates and 67 experts.

2.2 Analysis Methods and Algorithms

The time-agnostic velocity gain factor γ has shown promising results. In [5], this feature was used along with the two-third power law and the Isogony Principle to relate pencil tip velocity to the radius of curvature of 2D shapes sketched by a subject as

$$v(t) = \gamma(t)k(t)^{1/3}$$

where v is the velocity of the tool tip and k is the euclidean curvature (i.e. the instantaneous radius of curvature of the tool path.)

The mean and standard deviation of γ for the left and right hand over the course of the training run was taken to create the 4 features for each trial $\sigma(\gamma_L), \mu(\gamma_L), \sigma(\gamma_R)$ and $\mu(\gamma_R)$. These features were used to train a state vector machine (SVM) to predict the skill level of each trial.

The accuracy of the model was evaluated based on its agreement with the FLS classification for the trial, i.e. whether each model correctly classified the trial's FLS score grouping. Trials were grouped as either novice, intermediate, or expert. The FLS score groupings are used as the ground truth data.

To test the expected accuracy of an SVM which uses $\sigma(\gamma_L), \mu(\gamma_L), \sigma(\gamma_R)$ and/or $\mu(\gamma_R)$ as features, a 10-fold cross validation was performed. Cross validation helps ensure that the data used for testing was not included in the model training, and thus did not bias the model. The model's score grouping prediction was compared to the FLS score grouping from the EDGE data set. Each fold of the 10-fold cross validation samples trials evenly from the three skill levels. In addition, the 10-fold cross validation was performed 10 times (creating new 10-fold sets each time) in order to average out any fluctuations in accuracy due to the particular samples chosen for each 10-fold. This means, 100 different models comprised of 100 different partitions of the data were generated, and the statistics for the accuracies of each model are communicated in the box plot figures. In this study, we were only testing discrimination between novice and expert skill levels.

In each of the box plots, the 25th and 75th percentiles are displayed as the box boundaries while the median is the central line in each box. The whiskers mark the most extreme non-outlier points, and the + points are outliers.

3 Results

Task specific models were generated as a basis for comparison to other task-specific methods. In Figure 1, the classification

Seattle Children's — Internal

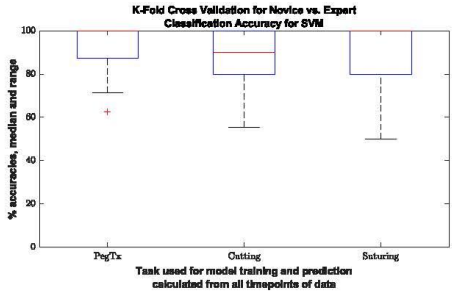


FIGURE 1: Prediction accuracy results of two-class classification, task specific trained model.

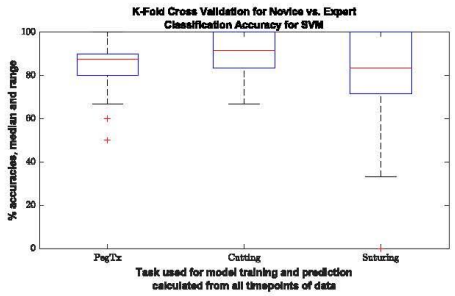


FIGURE 2: Prediction accuracy results of two-class classification, model trained in task-agnostic manner.

accuracies of task-specific models are displayed. To generate this figure, trials were separated based on which task was performed and a model to represent skill for each specific task was generated. A 10-fold cross-validation was done on each of these 3 partitions. This figure shows that for PegTx and Suturing models had a median classification accuracy of 100% per model.

In Figure 2, one model was trained using the full dataset in a task-agnostic manner (using trials across all tasks), and a similar cross validation technique to Figure 1 was performed. This figure shows good prediction accuracy for all three tasks despite the fact that the model was trained task-agnostic.

In Figure 3, five different models were trained using the full dataset in a task-agnostic manner, four of which were trained based on a single feature. The "all" column of this figure uses the same model from Figure 2. This shows the relative strength of each feature in prediction accuracy, where the accuracy of models trained using $\sigma(\gamma_R)$ or $\sigma(\gamma_L)$ alone are close in accuracy to the models trained using all four features.

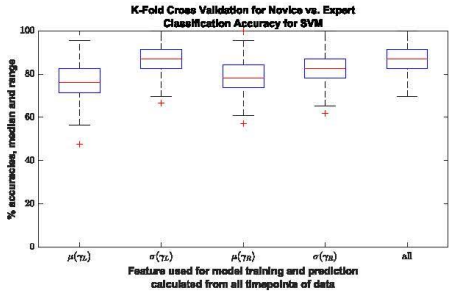


FIGURE 3: Prediction accuracy results of two-class classification, models trained using either single or all four features.

4 Interpretation

It was shown that task and time agnostic isogony-based features can be used to train automated skill evaluation models with good agreement to rough groupings of FLS scores. The ability to automate skill evaluation will allow surgeons in training to obtain feedback faster and more frequently. Further work will incorporate more trials to train more consistent models. These γ features will also be applied to other machine learning algorithms and used for 3-class classification to discern all FLS skill levels.

REFERENCES

[1] Kowalewski, T. M., 2012. "Real-time quantitative assessment of surgical skill". PhD thesis, University of Washington.

[2] Kowalewski, T. M., Comstock, B., Sweet, R., Schaffhausen, C., Menhadji, A., Averch, T., Box, G., Brand, T., Ferrandino, M., Kaouk, J., et al., 2016. "Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills tasks". The Journal of urology, **195**(6), pp. 1859–1865.

[3] Kowalewski, T. M., White, L. W., Lendvay, T. S., Jiang, I. S., Sweet, R., Wright, A., Hannaford, B., and Sinanan, M. N., 2014. "Beyond task time: automated measurement augments fundamentals of laparoscopic skills methodology". Journal of Surgical Research, **192**(2), pp. 329–338.

[4] Rana, J., and Kowalewski, T., 2014. "Feasibility of a low-cost instrumented trocar for universal surgical procedure analyses". Journal of Medical Devices, **8**(3), p. 030936.

[5] Lacquaniti, F., Terzuolo, C., and Viviani, P., 1983. "The law relating the kinematic and figural aspects of drawing movements". Acta psychologica, **54**(1-3), pp. 115–130.

Seattle Children's — Internal

Appendix D

Int J CARS manuscript No.
(will be inserted by the editor)

The Effect of Video Playback Speed on Surgeon Technical Skill Perception

Jason D. Kelly · Ashley Petersen Ph.D. · Thomas S. Lendvay M.D. · Timothy M. Kowalewski Ph.D.

Received: date / Accepted: date

arXiv:1911.00737v1 [cs.HC] 2 Nov 2019

Abstract *Purpose:* Finding effective methods of discriminating surgeon technical skill has proven a complex problem to solve computationally. Previous research has shown that obtaining non-expert crowd evaluations of surgical performances is as accurate as the gold standard, expert surgeon review [1]. The aim of this research is to learn whether crowd- sourced evaluators give higher ratings of technical skill to video of performances with in- creased playback speed, its effect in discriminating skill levels, and whether this increase is related to the evaluator consciously being aware that the video is being manually manipulated.

Methods: A set of ten peg transfer videos (5 novices, 5 experts), were used to evaluate the perceived technical skill of the performers at each video playback speed used (0.4x-3.6x). Objective metrics used for measuring technical skill were also computed for comparison by manipulating the corresponding kinematic data of each performance. Two videos of an expert and novice performing dry lab laparoscopic trials of peg transfer tasks were used to obtain evaluations at each playback speed (0.2x-3.0x) of perception of whether a video is played at real-time playback speed or not.

Results: We found that while both novices and experts had increased perceived technical skill as the video playback was increased, the amount of increase was significantly greater for experts. Each increase in the playback speed by 0.4x was associated with, on average, a 0.72-point increase in the GOALS score (95% CI: 0.60-0.84 point increase; $p < 0.001$) for expert videos and only a 0.24-point increase in the GOALS score (95% CI: 0.13-0.36 point increase; $p < 0.001$) for novice videos.

Conclusion: Due to the differential increase in perceived technical skill due to increased playback speed for experts, the difference between novice and expert skill levels of surgical performances may be more

Jason Kelly
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN E-mail: kell1917@umn.edu

A. Petersen
Division of Biostatistics, University of Minnesota, Minneapolis, MN

T. Lendvay
Department of Urology, Seattle Children's Hospital, Seattle, WA

T. Kowalewski
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN

Seattle Children's — Internal

easily discerned by manually increasing the video playback speed.

Keywords Crowd Sourcing Video Playback Surgical Technical Skill Speed Perception Bias

1 Introduction

Medical errors make up a third of all deaths in the United States, one of the largest contributors of which are surgical errors [2]. Technical surgical skill is directly related to patient outcomes [3], but it remains a difficult computational task to correctly classify surgeons into skill levels with a compelling level of accuracy, i.e. never misclassifying an 'obvious novice' as an 'obvious expert' and vice versa - the MAC Criterion [4]. The de facto gold standard for evaluating technical skill is video evaluation by an expert surgeon using Likert-scale assessment metrics, in which evaluators submit ratings on an anchored scale of 1-5. Using crowds of *non-expert* evaluators are a surprisingly accurate way to inexpensively and rapidly obtain skill level ratings for videos of surgical performances, with a pass/fail rate capable of matching 100% of ratings by expert surgeons [5]. The fact remains, however, that humans can be biased in their thinking, and subjective metrics of rating performances can lead to results we would not expect from computational models of evaluation.

One of the most popular laparoscopic surgical skill assessment metrics is the Global Operative Assessment of Laparoscopic Skills (GOALS), which is the most common objective assessment tool for laparoscopic performance studies [6], [7]. The subdomains in this met- ric include: bimanual dexterity, tissue handling, efficiency, depth perception, and autonomy. Task time is not a direct metric used to evaluate the technical skill of laparoscopic surgeons with tools like GOALS, however time for task completion is often seen as one of the most predictive objective forms of evaluating technical skill, as seen in Fig. 1 [8], [9].

1.1 Biological Motion Perception

Previous research involving point light walkers (PLW), 10-12 dots of light illuminated on a screen to correspond to the human joints in walking motion, (illustrated in Fig. 2), revealed that humans are excellent at recognizing more subtle qualitative characteristics like the gen- der or emotion of a PLW by inferring the gait from a few moving points [10]. Once the PLW's walking gait speed is changed however, the ability to recognize gender is diminished [11]. This and other work suggest that biological motion perception is somehow tuned to the speed of the motion involved [12].

Due to task time being such a strong indicator of surgical technical skill, we speculate about whether a video artificially sped up to simulate a quicker time for task completion will lead to an altered perception of technical skill, or if human raters evaluate other, more nuanced, qualities of the performance.

1.2 Objective Metrics

Other metrics exist for measuring technical skill, with varying degrees of success at classifying novice and expert skill levels. These techniques utilize kinematic tool tip data, as opposed to using video

Seattle Children's — Internal

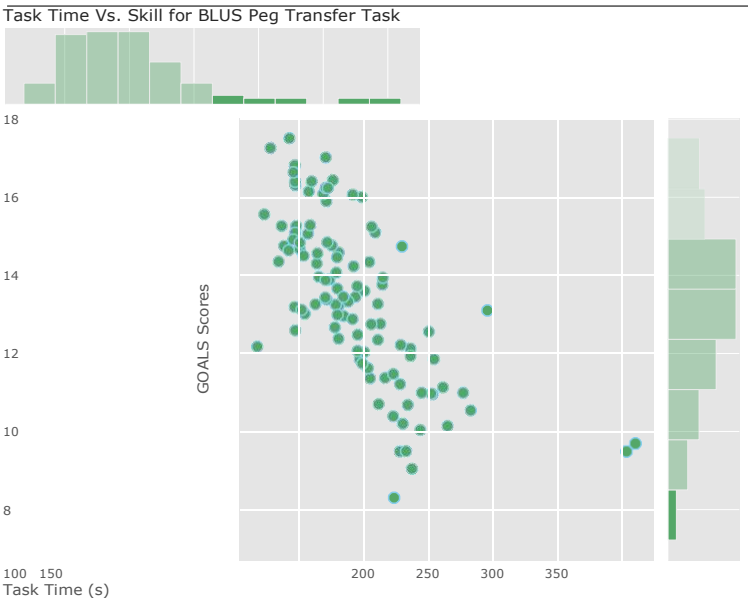


Fig. 1: Task time completion vs. GOALS scores for all peg transfer performances in the Basic Laparoscopic Urologic Skills (BLUS) dataset [13].

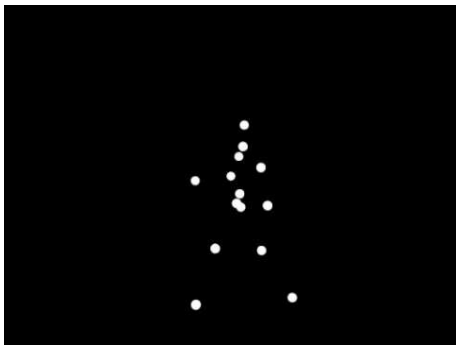


Fig. 2: A point light walker, used in biological motion research. Despite highly impoverished data (points only, no video), humans seem to quickly infer human gait within a few steps of animation and even extract subtle cues like gender or emotional state [10], [11].

frames to make measurements. Metrics which use algorithms or computational methods to evaluate performance are sometimes referred to as Automated Performance Metrics (APMs) [9]. The simplest metric to measure, and often the most ac- curate, is time of task completion. The following APMs seek to compute information more complex than speed to account for technical skill, by conversely using the form of movements made.

Prior work has discovered that recovering stroke patients have a decrease in 'jerky' non- smooth movements as they progress in rehabilitation therapy [14]. The jerk cost of all tool movements performed in a task has also been used to measure accuracy and skill in areas in which skill is needed [15]. The jerk cost is computed by taking the integral of the squared jerk summed for both the left and right hand over the course of the performance, shown in Equation 1, in which T is total time of task performance and x is the magnitude of the movement.

$$\text{Jerk} \quad \int_0^T \left| \dddot{x}(t) \right|^2 dt \qquad (1)$$

Spectral Arc length (SAL) is also related to smoothness of a movement and aims to use the tangential velocity of hand movements to compute the overall smoothness of a task. This has been used to

Seattle Children's — Internal

differentiate skill levels of medical professionals in the past [16], [17]. To compute the SAL, the motion in a task must be segmented into specific hand grasping movements, with the corresponding speeds and durations known for each grasp. The Fourier magnitude spectrum transform of the speed profile is then computed and normalized with respect to its zero-frequency value. The smoothness is then measured for each segmented grasp, shown in Equation 1.2, in which ω is frequency, ω_c is cutoff frequency, and V is the Fourier magnitude spectrum of speed. Finally, a weighted average is taken across all grasping motions to arrive at an SAL metric for the task as a whole.

$$SAL = -\frac{1}{\omega_c}\int_0^{\omega_c}\left(\frac{d\hat{V}(\omega)}{d\omega}\right)^2 d\omega \qquad \hat{V}(\omega) = \frac{V(\omega)}{V(0)} \tag{2}$$

Instead of relying on the form of a grasping movement, taking the sum of all movements which occur is an additional metric that can be used to evaluate skill in some domains in which it is believed that novice performers on average perform more grasping movements than experts. A movement can be calculated in several ways, but the most common way of recording is to define a threshold of speed at which when the performer's speed falls below that threshold, a movement has ended and a new grasping motion starts once the magnitude of the velocity has risen above that threshold again. Counting each time this threshold is passed gives this total count of all movements conducted during the task [18].

All of these APMs which may be used to compute the skill of a surgeon have in common that once the speed playback of the data is manipulated, the results normally do not lead to any more separation between skill levels, as changing the magnitude of speed will not affect calculations to a large degree.

Our motivation in this work is to learn whether, by increasing video playback speed, novices and experts will be perceived by crowds as more skilled when they appear to be moving faster, *and* if experts and novices will have different rates of perceived change as the playback speed is increased. We will do this by evaluating the ability to discriminate obvious novices from experts of both human raters and popular tool motion metrics (APMs), at different playback speeds. We expect each APM's ability to discriminate skill to operate as the experimental control in this study, with a negligible change in separation between skill levels as speed is increased. Finally, we seek to investigate how an evaluator's likely conscious awareness of whether a video is sped up or slowed down relates to their perception of skill.

Table 1: Likert-scale technical skill perception questionnaire for a manually sped-up video, from four domains of the GOALS assessment metric.

Seattle Children's — Internal

Score	**Depth Perception**
(1)	Constantly overshoots target, wide swings, slow to correct
(2)	
(3)	Some overshooting or missing of target, but quick to correct
(4)	
(5)	Accurately directs instruments in the correct plane to target

Score	**Bimanual Dexterity**
(1)	Uses only one hand, ignores non-dominant hand, poor coordination between hands
(2)	
(3)	Uses both hands, but does not optimize interaction between hands
(4)	
(5)	Expertly uses both hands in a complementary manner to provide optimal exposure

Score	**Efficiency**
(1)	Uncertain, inefficient efforts; many tentative movements; constantly changing focus or persisting without progress
(2)	
(3)	Slow, but planned movements are reasonably organized
(4)	
(5)	Confident, efficient and safe conduct, maintains focus on task until it is better performed by way of an alternative approach

Score	**Tissue Handling**
(1)	Definitely sped-up video
(2)	
(3)	Handles tissue reasonably well, minor trauma to adjacent tissue (i.e. occasional unnecessary bleeding or slipping of the grasper)
(4)	
(5)	Handles tissues well, applies appropriate traction, negligible injury to adjacent structures

Table 2: Likert-scale speed perception questionnaire for a manually sped-up video.

Score	**Does this video appear to have been altered to have an increased playback speed?**
(1)	Definitely sped-up video
(2)	Likely sped up video
(3)	Not sure / I don't know
(4)	Not likely sped up video
(5)	Definitely real-time playback speed (unaltered speed)

2 Methods

2.1 Dataset

This study used the Basic Laparoscopic Urologic Study (BLUS) dataset, described in detail in [1]. This dataset arose from a gap in the field, in which no educational surgical certification process existed for urologic surgery, as opposed to how the Fundamentals of Laparoscopic Surgery (FLS) exists for general surgical procedures [19], [20], [21]. The BLUS training curriculum aimed to address urology appropriate skills improvement by recording video performances in an initial validation project of over 450 videos [13].

This dataset contains 454 videos of surgical performances consisting of four surgical tasks (110 peg transfer, 110 pattern cutting, 115 suturing, 119 clip applying), which are performed by medical students, urology residents, fellows and faculty surgeons from eight academic urology training centers in the United States [15]. Each trial of a surgeon per- forming one of the four tasks was recorded at 30 fps with a fixed camera-position of the laparoscopic tools interacting with the training field. Each trial additionally has kinematic data, sampled at 30 Hz, logging the tooltip positions, grasping force, and the jaw angles during the performance, as well as demographic information for each performer being obtained. A GOALS score for each performance was also previously obtained from either expert or crowd evaluation.

Seattle Children's — Internal

In previous research the Peg Transfer task has been shown to be one of the most easily differentiable tasks for surgical technical skill. Although the Peg Transfer task is criticized as the least clinically relevant, its clear ability to separate novice/expert skill levels was preferred to explore the research questions herein. Ten videos from the peg transfer task were used, in which five 'obvious experts' and five 'obvious novices' were chosen as baseline definitions of skill [4]. Here an obvious novice is defined as someone who should never be allowed to operate and obvious experts as surgeons who should never be disqualified from operating. An obvious expert was chosen such that the performer was in the top 20% of experience levels (attending surgeon or faculty urologist), previously obtained GOALS scores, and task completion times of all peg transfer tasks. The obvious novices were chosen in the same fashion such that they were in the bottom quintile of these domains. This method aimed to provide two well discriminated clusters of skill levels.

Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each non-expert crowd worker was paid an average of $0.10 to watch and evaluate a video, depending on the video duration. A web domain was created for which Turkers would be redirected to, where they submitted a consent form and were asked questions about videos. Two different kinds of experiments were conducted: *technical skill perception* and *speed perception*.

2.2 Experiment 1: Technical Skill Perception

Technical skill perception was measured by surveying non-expert crowds to give each video performance a GOALS score by rating each of the 4 subdomains shown in Table 1. Forty "turkers" were recruited per video, in which each video at each playback speed was inde- pendently submitted to the website in order to avoid a grouping bias. Videos were altered to speeds in the range 0.4x-3.6x, moving in intervals of 0.4, edited using FFmpeg [22], in which frames were either taken out or added in order to create the resulting playback speed. The score from each of four subdomains (Depth Perception, Bimanual Dexterity, Efficiency, and Tissue Handling) were summed to create a cumulative score for each performance in the range of 4-20. The mean of each video's cumulative GOALS score was recorded, with a 95% confidence interval.

2.3 Experiment 2: Speed Perception

The video playback speed perception test was conducted by using one novice and one expert from the peg transfer task and asking the evaluator if they thought the video was being manually edited. If we showed them a video with a decreased playback speed, the worker was asked whether they thought the video was running at real-time speed or was manually edited to be slowed down. In the same way, workers were asked if the video was being sped up or played at real-time playback speed for videos which had an increased video playback speed. These Likert-scale questions are shown in Table 2 for a video in which the playback speed was increased. Speeds tested ranged from 0.2x-3.0x, moving in intervals of 0.2x to verify the range of speeds to use for the technical skill perception tests.

2.4 APM Validation at Each Playback Speed

The APMs mentioned in Section 1.2 were also recorded at these various playback speeds for comparison to crowd evaluations. These were recalculated by scaling the magnitude of the speeds in the kinematic data to be equivalent to the speed displayed in a given video playback speed, then computing these objective metrics. For the jerk cost, this involved using a Holoborodko smooth noise-robust differentiator [23]. For the others this involved simply scaling the magnitude of speed and inputting the result to the pre-created function for each corresponding APM. All APMs were computed in Python and MATLAB, [24], [25].

Seattle Children's — Internal

3 Results

3.1 Technical Skill Perception

The mean of the GOALS score for each obvious expert and novice are shown in Fig. 3. For expert videos, each increase in the playback speed by 0.4x was associated with, on average, a 0.72-point increase in the GOALS score (95% CI: 0.60-0.84 point increase; $p < 0.001$). On average these scores appear to increase within a sublevel of the playback speeds around 0.5x to 2.4x, and then level out at all remaining playback speeds. For novice videos, each increase in the playback speed by 0.4x was associated with, on average, a 0.24-point increase in the GOALS score (95% CI: 0.13-0.36 point increase; $p < 0.001$). Thus, while both experts and novices had increased perceived technical skill as the playback speed was increased, the gain was significantly greater for experts. The experts had, on average, a 0.47-point greater increase (95% CI: 0.31-0.64; $p < 0.001$) in the GOALS score, compared to novices, for each 0.4x increase in playback speed. Fig. 4 shows the increase in the efficiency subdomain as well as the average of the other three domains to visualize whether efficiency (seen as the most related to speed) is the only increasing domain. As shown, it is clear that the other domains increased at nearly the same rate.

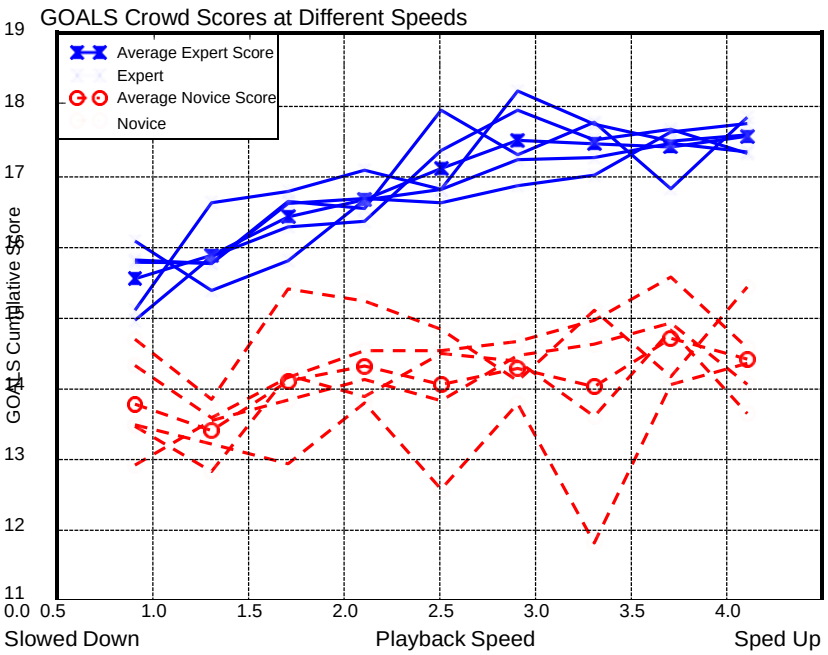


Fig. 3: All mean crowd evaluations from each novice and expert peg transfer task at various video playback speeds. (Each solid marker indicates N = 40).

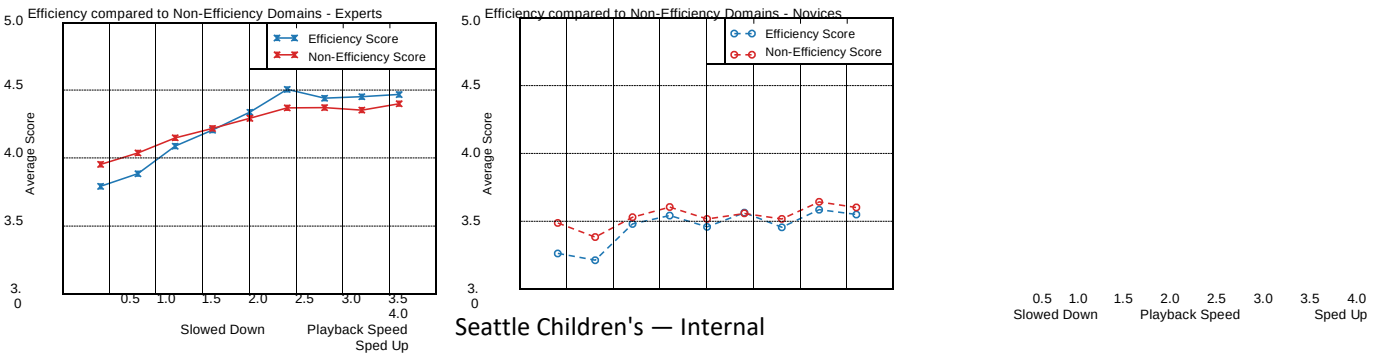


Fig. 4: The efficiency subdomain as compared to the average of the other GOALS subdomains, for experts and novices.

The APMs which were calculated with the manipulated kinematic data are shown in Fig.5. As shown, most metrics show very little change at each of the different speeds, but even if they change, performers' technical skills are not able to be discriminated between expert and novice skill any more easily. These metrics were also calculated by incorporating a combination of scaling the speeds and removing indices corresponding to the frames which would be removed from videos when the playback speed is increased. These figures aren't included, but the results exhibit the same lack of separation between the skill groups at various playback speeds. Fig. 6 shows the difference in mean for experts and novices, tested with both the crowd scores and the highest performing APM, spectral arc length, to illustrate the difference in skill discrimination between crowd workers and the most accurate APM as speed is increased.

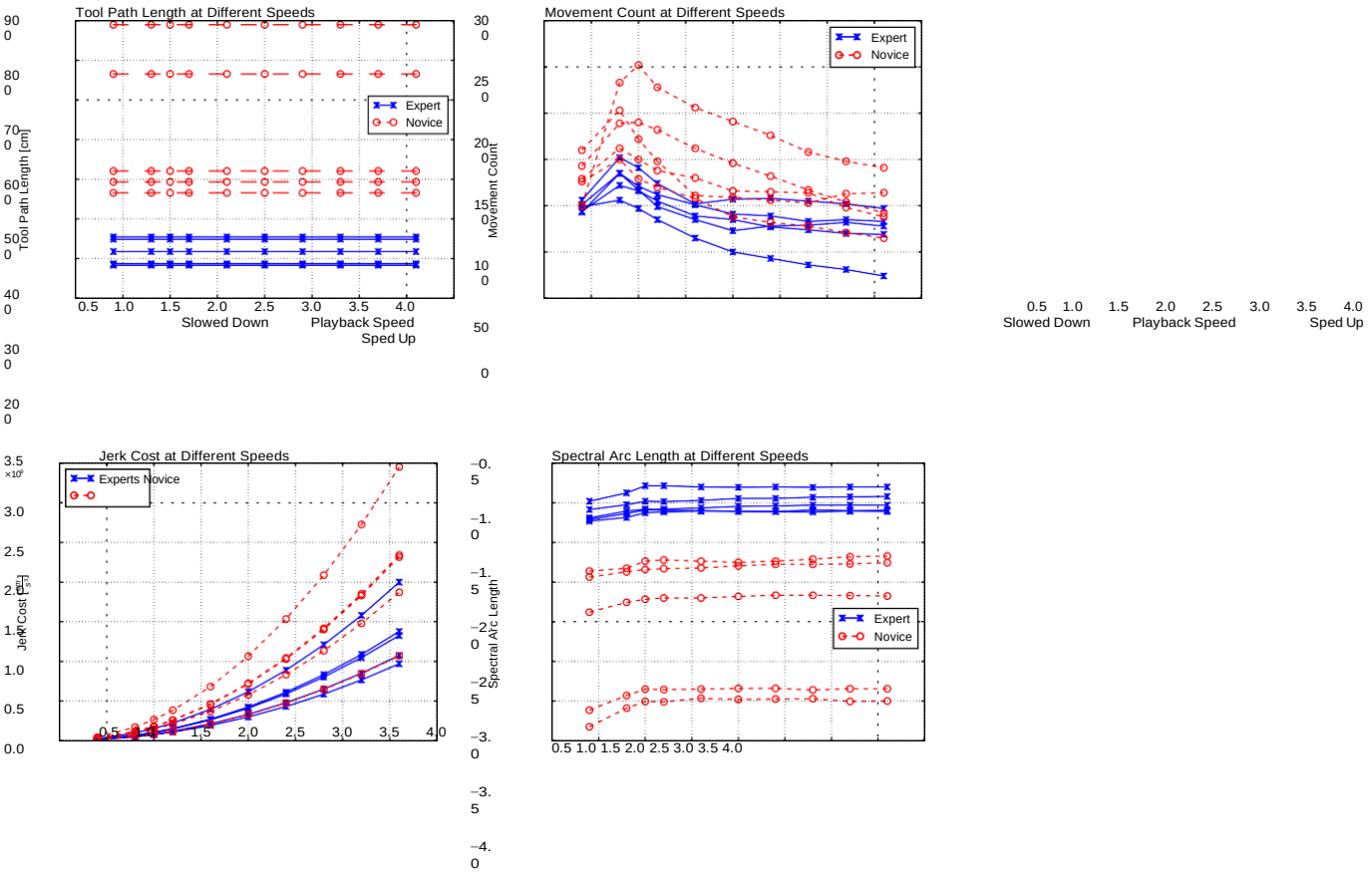


Fig. 5: Objective technical skill metrics for the obvious experts and novices across various speeds.

3.2 Playback Speed Perception

The speed perception results for an obvious novice and obvious expert are plotted in Fig. 7a. The scores were obtained by treating each of the scores in Table 2 as a score, and obtaining the mean score for each video evaluated, such that a higher score leads to a higher perception of the video being played at a real-time playback speed. Both novices and experts displayed similar perceptions of the video either having an increased or decreased video playback speed. In Fig. 7b, a two sample independent t-test was computed between the two groups at each of the playback speeds, to visualize the difference in mean speed perception

between experts and novices. All of the speeds show no significant difference in playback speed recognition. The perceived video playback speed scores are compared with the same video's crowd scores for the GOALS assessment metric in Fig. 8.

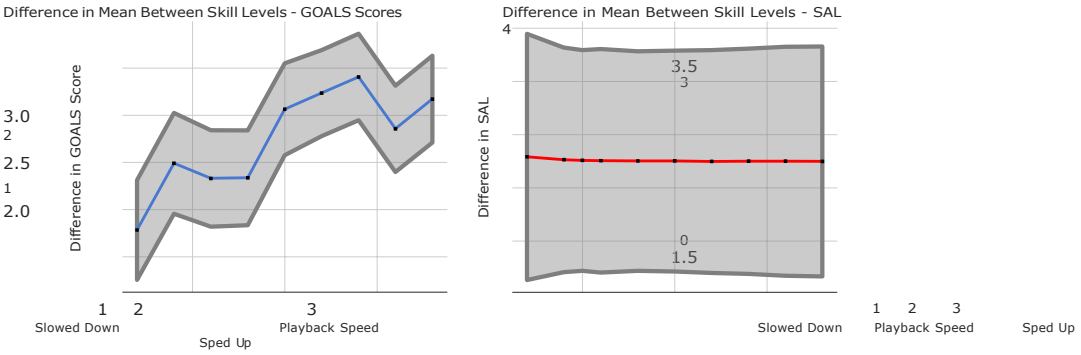


Fig. 6: Difference in mean between skill levels from the highest performing APM and from crowd scores, with a 95% CI in the shaded regions.

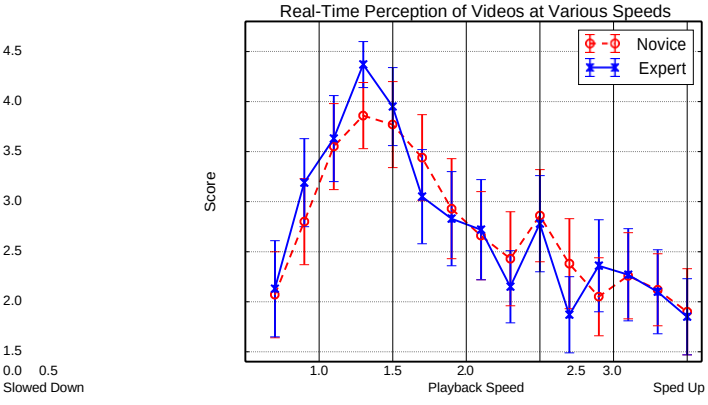
4 Conclusion

The results from the technical skill perception study give support to our initial hypothesis that increasing the video playback speed would increase the ratings of surgical performances. Surprisingly, however, we discovered that novice performances receive a much lower increase in score, which is almost negligible. This finding elucidates the notion that despite increasing the video playback speed of a novice performance, non-expert crowd workers are still able to spot the more obvious mistakes made by these novices. How-ever, it appears that crowd evaluators are biased to speed when they evaluate expert performances. This could be due to expert performers appearing as though their movements are even smoother at quicker playback speeds and the few mistakes they are making being 'washed out' or not emphasized at quicker playback speeds. With 135,000 surgeons in the U.S. and the growing need to objectively quantify surgical skills to ensure public safety, methods to rapidly triage technical skill are required. Our observations may provide an easier and cheaper way of discriminating novices from experts than with using expert surgeon evaluation.

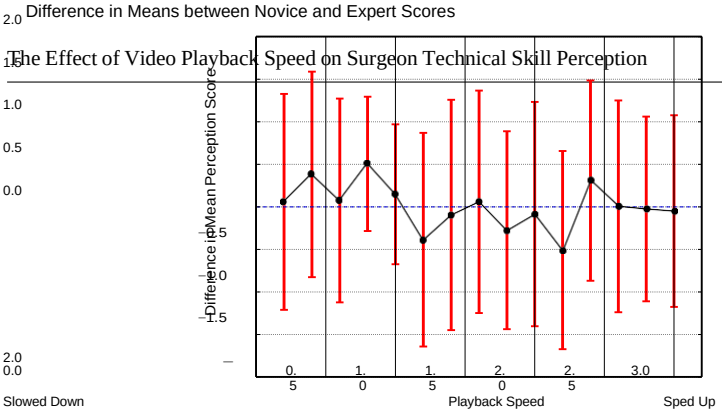
To illustrate whether the 'Efficiency' domain of the GOALS assessments was the only domain increasing, as this domain is most closely related to speed, a scatter plot for the average efficiency domain and average of the other three domains for novices and experts is shown in Fig. 4. The novices clearly show no difference in Efficiency vs. the other three domains. In addition, the three remaining domains for experts appear to increase at almost the same rate as the Efficiency domain. It appears that crowds are also biased to give higher ratings in other domains which aren't necessarily related to speed.

The results from the real-time playback perception study validate our initial hypothesis by showing us that crowds are able to accurately discern when a video has been manually edited to have a different playback speed. Although, as shown in Fig. 7a, there is apparently no difference in perception between an obvious novice performance and an obvious expert performance.

As shown in our comparison to objective metrics, no other used method of objectively obtaining technical skills scores will show the same amount of improvement for increases in playback speed. There may be information that humans can decipher which objective metrics or machine learning algorithms cannot.



(a) All mean crowd evaluations of video speed perception which ends in real-time playback speeds. (Each solid marker indicates N = 40).



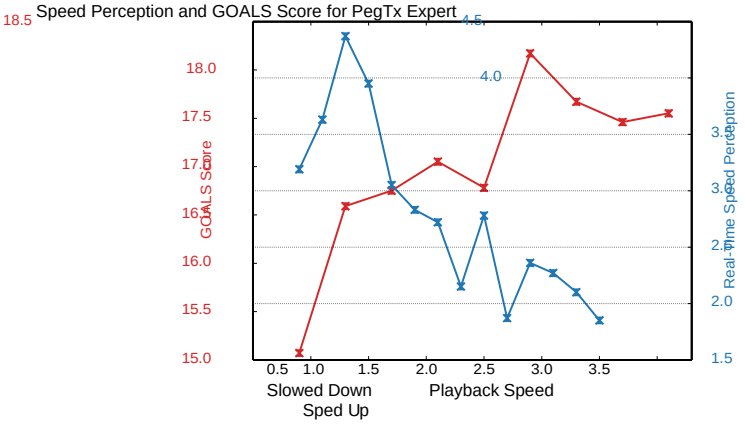
(b) Difference in Mean between Expert and Novice evaluations.

Fig. 7: Comparison of Novice and Expert Speed Perception at each playback speed. (There were $N = 40$ *unique human evaluations/skill level/playback speed.* The error bars represent 95% confidence intervals.)

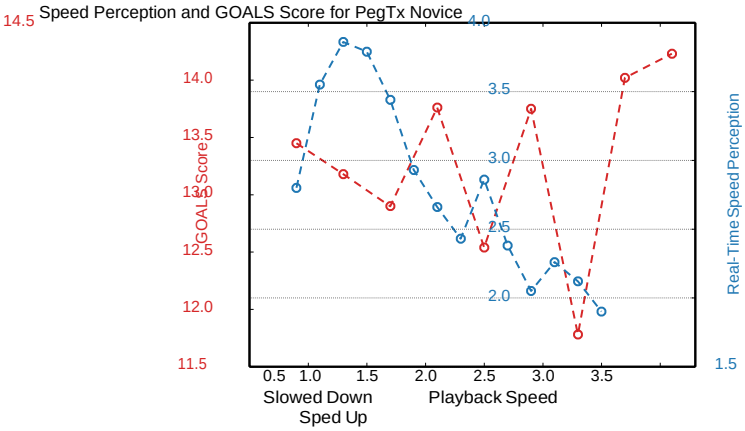
We conclude that increasing the video playback speed of performances of dry lab laparoscopic training tasks could provide a cheap and easy way to discriminate experts from novices as the separation in GOALS scores between novices and experts appears to increase as video playback is increased. A limitation of this study is that we only sampled ten videos of laparoscopic training procedures. Additional investigation with a larger dataset of more clinically relevant performances is required to conclude whether this observation extends to actual surgical case evaluation.

Compliance with Ethical Standards

Conflicts of Interest The authors declare that they have no conflict of interest.



(a) Expert



(b) Novice

Fig. 8: Speed Perception compared with technical skill perceived at various speeds.

Ethical Standard All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research commit- tee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Funding This work was supported, in part, by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-15-2-0030, the National Science Foundation CAREER grant under Award No. 1847610, as well as the National Institutes of Health's National Center for Advancing Translational Sciences, grant UL1TR002494. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense, the National Science Foundation, or the National Institutes of Health's National Center for Advancing Translational Sciences.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Seattle Children's — Internal

References

1. Kowalewski T, Comstock B, Sweet R, Schaffhausen C, Menhadji A, Averch T, Box G, Brand T, Fer- randino M, Kaouk J, Knudsen B, Landman J, Lee B, Schwartz BF, McDougall E, and Lendvay TS (2015) Crowd-Sourced Assessment of Technical Skills for Validation of Basic Laparoscopic Urologic Skills (BLUS) Tasks. The Journal of Urology 195(6):1859–1865.
2. Makary M and Daniel M (2016) Medical error - the third leading cause of death in the US. The BMJ.
3. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, and Birkmeyer NJ (2013) Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine 369(15):1434–1442.
4. Dockter R, Lendvay TS, Sweet RM, and Kowalewski TM (2017) The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. The International Journal of Computer-Assisted Radiology and Surgery 12:1151–1159.
5. Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B, Kuksenok K, Aragon C, Holst D, and Lendvay T (2013) Crowd-Sourced Assessment of Technical Skills: a Novel Method to Evaluate Surgical Performance. Journal of Surgical Research 187(1):65–71.
6. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondre K, Stanbridge D, and Fried GM (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. The American Journal of Surgery 190(1):107–113.
7. Gumbs AA, Hogle NJ, and Fowler DL (2007) Evaluation of Resident Laparoscopic Performance Using Global Operative Assessment of Laparoscopic Skills. Journal of the American College of Surgeons 204(2):308–313.
8. Kowalewski TM, White LW, Lendvay TS, Jiang IS, Sweet RS, Wright A, Hannaford B, and Sinanan MN (2014) Beyond task time: automated measurements augments fundamentals of laparoscopic skills methodology. Journal of Surgical Research 192(2): 329–338.
9. Hung A, Chen J, Che Z, Nilanon T, Jarc A, Titus M, Oh PJ, Gill IS, and Liu Y (2018) Utilizing Machine Learning and Automated Performance Metrics to Evaluate Robot-Assisted Radical Prostatectomy Performance and Predict Outcomes. Journal of Endourology 32(5).
10. Barclay CD, Cutting JE, and Kozlowski LT (1978) Temporal and spatial factors in gait perception that influence gender recognition. Percpetion and Psychophysics 23(2):145–152.
11. Veto P, Einhauser W, and Troje NF (2017) Biological motion distorts size perception. Scientific Reports 7(42576):1–6.
12. Jacobs A, Pinto J, and Shiffrar M (2004) Experience context and the visual perception of human movement. Journal of Experimental Psychology 30(5):822–835.
13. Seete RM, Beach R, Sainfort F, Gupta P, Reihsen T, Poniatowski LH, and McDougall EM (2012) In- troduction and validation of the American Urological Association Basic Laparoscopic Urology Surgery skills curriculum. Journal of Endourology, 26: 190.
14. Flash T and Hogan N (1985) The Coordination of Arm Movements: An Experimentally Confirmed Mathematical Model. The Journal of Neuroscience 5(7):1688-1703.
15. Kowalewski TM, Seet R, Lendvay TS, Menhadji A, Averch T, Box G, Brand T, Ferrandino M, Kaouk J, Knudsen B, Landman J, Lee B, Schwartz BF, and McDougall E (2016) Validation of the AUA BLUS Tasks. Journal of Urology 195: 998.
16. Bajcsy A, Losey DP, O'Malley MK, and Dragan AD (2018) Learning from Physical Human Corrections, One Feature at a Time. In Proceedings of 2018 ACM/IEEE International Conference on Human-Robot Interaction, NY, USA.
17. Balasubramanian S, Melendez-Calderon A, and Burdett E (2012) A Robust and Sensitive Metric for Quantifying Movement Smoothness. IEEE Transactions on Biomedical Engineering 59(8):2126–2136.
18. Kowalewski TM (2012) Real-time quantitative assessment of surgical skill. PhD Thesis, University of Washington.
19. Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, and Meakins JL (1998) Development of a model for training and evaluation of laparoscopic skills. American Journal of Surgery 175: 482.
20. Fried GM (2008) FLS assessment of competency using simulated laparoscopic tasks. Journal of Gastroenterology Surgery 12: 210.
21. Peters JH, Fried GM, Swanstrom LL, Soper NJ, Silin LF, Schirmer B, and Hoffman K (2004) Develop- ment and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery, 135: 21.
22. FFmpeg Developers (2016) ffmpeg tool (Version 4.1.3) [Software]. Available form http://ffmpeg.org/.
23. Holoborodko P (2008) Smooth Noise Robust Differentiators. http://www.holoborodko.com/pavel/numerical- methods/numerical-derivative/smooth-low-noise-differentiators/
24. Python Software Foundation. Python Language Reference, version 3.6. Available at http://www.python.org

MATLAB R2019a, The Mathworks Inc. Natick, Massachusetts.

Appendix E

Int J CARS manuscript No.
(will be inserted by the editor)

Bidirectional Long Short-Term Memory for Surgical Skill Classification of Temporally Segmented Tasks

Jason D. Kelly · Ashley Petersen Ph.D. · Thomas S. Lendvay M.D. · Timothy M. Kowalewski Ph.D.

Received: date / Accepted: date

Abstract *Purpose:* The majority of historical surgical skill research typically analyzes holistic summary task-level metrics to create a skill classification for a performance. Recent advances in machine learning allow time series classification at the sub-task level, allowing predictions on segments of tasks, which could improve task level technical skill assessment.

Methods: A bidirectional long short-term memory (LSTM) network was used with 8 second windows of multidimensional time-series data from the Basic Laparoscopic Urologic Skills (BLUS) dataset. The network was trained on Experts and Novices from four common surgical tasks. Stratified cross validation with regularization was used to avoid overfitting. The misclassified cases were re-submitted for surgical technical skill assessment to crowds using Amazon Mechanical Turk to re-evaluate, to analyze the level of agreement with previous scores.

Results: Performance was best for the suturing task, with 96.88% accuracy at predicting whether a performance was an expert or novice, with 1 misclassification, when comparing to previously obtained crowd evaluations. When compared with expert surgeon ratings, the LSTM predictions resulted in a Spearman coefficient of 0.89 for suturing tasks. When crowds re-evaluated misclassified performances it was found that for all 5 misclassified cases from peg transfer and suturing tasks, the crowds agreed more with our LSTM model than with the previously obtained crowd scores.

Conclusion: This technique shows promise both for classifying performances of surgical tasks and evaluating sub-task level segments of performances. This method could be used for near real-time feedback to surgeons. Further work will be done to expand this methodology to a larger cohort of surgical performances including robot-assisted surgery.

Jason Kelly
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN
E-mail: kell1917@umn.edu

A. Petersen
Division of Biostatistics, University of Minnesota, Minneapolis, MN

T. Lendvay
Department of Urology, Seattle Children's Hospital, Seattle, WA

T. Kowalewski
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN

Seattle Children's — Internal

Appendix F

Int J CARS manuscript No.
(will be inserted by the editor)

The Effect of Video Playback Speed on Perception of Technical Skill in Robotic Surgery

Jason D. Kelly, Nicholas Heller, Ashley Petersen Ph.D., Thomas S., Lendvay M.D., Timothy M. Kowalewski Ph.D.

Received: date / Accepted: date

Abstract *Purpose:* Previous research has shown that obtaining non-expert crowd evaluations of surgical performances concords with the gold standard of expert surgeon review, and that faster playback speed increases ratings for videos of higher-skilled surgeons in laparoscopic simulation [1], [2]. The aim of this research is to extend this investigation to real surgeries that use non-expert crowd evaluations. We address two questions (1) whether crowds award more favorable ratings to videos shown at increased playback speeds, and (2) if crowd evaluations of the first minute of a surgical procedure differ from crowd evaluations of the entire performance.
Methods: A set of 56 videos of robotic surgeries (for 28 “experts”, and 28 “proficient” surgeons), were used to evaluate their perceived technical skill at five playback speeds for the first minute of the previously rated performance, using the Global Evaluative Assessment of Robotic Skills (GEARS) assessment. Each video was subsequently rated at 1x speed to obtain objective ratings.
Results: Crowds on average did rate videos higher as playback speed was increased, for both skill groups. Each increase in the playback speed was associated with an average 0.16- point increase in score for experts and 0.27-point increase for proficient surgeons, with both groups having roughly equal scores at the quickest speed. Additionally, 11 of the 28 surgeons in both groups were rated as belonging to the opposing category when viewing only the first minute of performance.
Conclusion: The observed increase in skill ratings with video playback speed replicates findings for laparoscopic experts in [2], and extends to the context of real robotic surgeries.

Jason Kelly
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN E-mail: kell1917@umn.edu

N. Heller
Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN

A. Petersen
Division of Biostatistics, University of Minnesota, Minneapolis, MN

T. LendvayDepartment of Urology, University of Washington, Seattle, WA

T. Kowalewski
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN

The results show that crowds do seem biased in rating surgeons as more highly skilled when they appear quicker. The large differences in skill labels when comparing the first minute of surgery to the entire 15 minute video warrants further investigation into how much perceived skill ratings vary in time.

Keywords Crowd Sourcing Video Playback Surgical Technical Skill Speed Perception Bias

1 Introduction

A third of all deaths in the United States are caused by medical errors, and surgical errors are one of the largest contributors to this [3]. Technical surgical skill is directly related to patient outcomes [4], but it

Seattle Children's – Internal

remains a difficult computational task to correctly classify surgeons into skill levels with a compelling level of accuracy, i.e. never misclassifying an ‘obvious novice’ as an ‘obvious expert’ and vice versa - the MAC Criterion [5]. The de facto gold standard for evaluating technical skill is video evaluation by an expert surgeon using Likert- scale assessment metrics, in which evaluators submit ratings on an anchored scale of 1-5. Using crowds of *non-expert* evaluators are a surprisingly accurate way to inexpensively and rapidly obtain skill level ratings for videos of surgical performances, with a pass/fail rating that matches 100% of pass/fail ratings by expert surgeons [6]. The fact remains, however, that humans can be biased in their thinking, and subjective metrics of rating performances can lead to results we would not expect from computational models of evaluation.

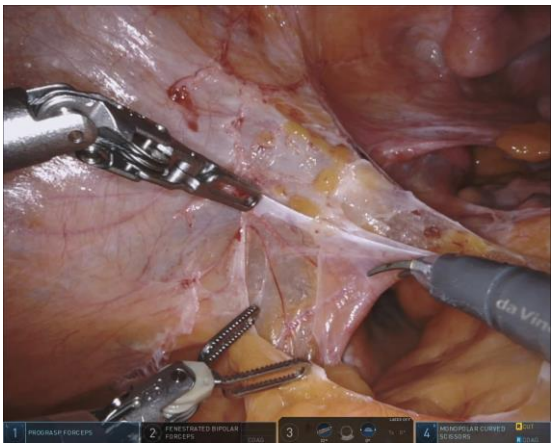


Fig. 1: Frame from a video collected for this study, recorded from a daVinci surgical robot, Intuitive (Sunnyvale, CA).

A popular laparoscopic surgical skill assessment metric is the Global Evaluative Assessment of Robotic Skills (GEARS), which is the most common assessment tool for robotic surgery skills [7]. The subdomains in this metric include: bimanual dexterity, efficiency, depth perception, force sensitivity and robotic control. Task time is not a direct metric used to evaluate the technical skill of laparoscopic surgeons with tools like GEARS, however time for task completion is often seen as one of the most predictive objective forms of evaluating

technical skill [8], [9]. However, there must be ways of objectively evaluating skill between multiple performances which were completed in the same time span.

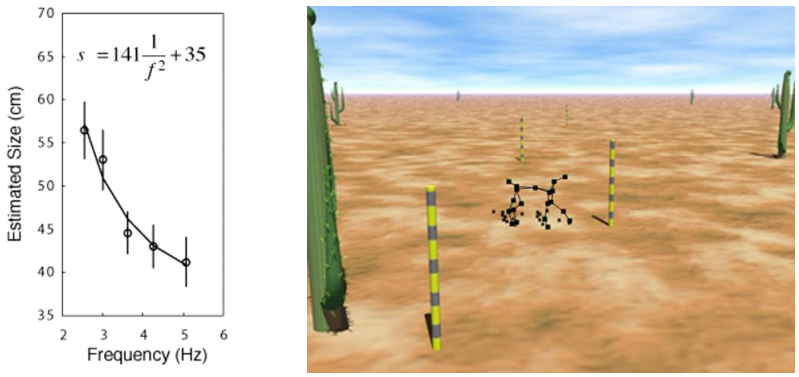


Fig. 2: Illustration of research results (left) which found the frequency of gait in simulated animals (right) affects the estimated size [10].

1.1 Changes in Perception due to Speed

Research in biological motion processing has observed a relationship between the gait frequency and estimated size of animals and other objects which replicate the joints of the human body. It was found that when a computer simulation was used to artificially modify the gait speed of these objects to unnaturally high speeds, study participants perceived the size of the moving objects as changing. As the gait frequency was increased, the estimated size of the object enlarged, shown in Fig. 2, [10], [11]. The results from this study may sug- gest a link between how humans evaluate biological motion and the speed of a movement. It is possible this phenomenon extends to other areas of evaluation, such as technical skill in surgery.

1.2 Technical Skill Ranges

Previous work has shown that crowds are able to discriminate the levels of novice and expert surgeons and their scores change as playback speed changes - even into motions appearing artificially sped up [2]. This was a limited study using a small dataset of laparoscopic train- ing exercises from the Basic Laparoscopic Urologic Skills dataset [12], [13], with extremes of skill levels (e.g. “obvious novice” medical students with no suturing experience). To test whether this phenomenon relates to actual surgical procedures, a larger dataset must be used which incorporates videos of real surgical footage and more practical ranges of surgeon skill. However, surgeons which are allowed to safely operate on patients can’t ethically be “obvious” novice surgeons, according to the MAC Criterion [5]. This means the surgeons in a dataset of real surgical footage will comprise a smaller range of technical skill levels,

with all surgeons being more comparably rated. This necessitates the need to further separate experts into two groups: proficient and expert surgeons. Proficient surgeons as stated in
[14] are defined as a surgeon who is well advanced in any branch of knowledge or skill. We will adopt this as our term for surgeons who are in the bottom quintile of scores from our surgical videos, as they are still well above average skilled surgeons.

Our motivation in this work is to investigate whether, by increasing video playback speed in real robotic surgery, surgeons will be perceived by crowds as more skilled when they appear to be moving faster. It was additionally desired in this work to analyze a subset of skill from previous findings, by only analyzing surgeons with above average skill levels, to see if smaller differences exist within this population. We will do this by evaluating the ability of non-expert crowd workers to discriminate proficient surgeons from experts. Finally, we seek to examine the effect of video duration on ratings following [15].

2 Methods

2.1 Dataset

This study used a novel Robotic Surgery Readiness (RSR) Study dataset, which consists of 343 videos of live robotic surgeries, with matching kinematic data also recorded, though unused here. These surgeries were performed by attending surgeons and trainees in urology, gynecology, and general surgery at the University of Washington Medical Center and the Puget Sound Veterans Administration. Each video was manually edited to include roughly the first 15 minutes of surgical activity performed by the surgeon. There are a wide variety of surgical procedures recorded from this dataset including: prostatectomy, cystectomy, hysterectomy, partial nephrectomy, and sacrocolpopexy. An image of a frame from one of these surgeries is shown in Fig. 1. A GEARS score for each performance was also obtained from crowd evaluation.

The range of scores for the RSR videos was fairly small, with most lying between 20- 22 out of 25. This is not a setback however, as it was desired to analyze a subset of skill, those with above average levels of surgical technical skill, to find if there are smaller differences in this group, as well as if previous findings extend to live surgical footage. 28 performances from the top quintile of scored performances and 28 from the bottom quin- tile of performances were used and given labels of ‘expert’ and ‘proficient’, respectively, keeping in mind that almost all of the performances would have objectively been considered

Seattle Children's — Internal

expert-like for a larger distribution of scores. For semantic analysis, the first minute of surgical activity was extracted from each 12-15-minute video for analysis by crowds.

Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each non-expert crowd worker was paid an average of $0.40 to watch and evaluate a video. A web domain was created for which Turkers would be redirected to, where they submitted a consent form and were asked questions about videos. Two different kinds of experiments were conducted: *technical skill perception at different playback speeds* and *sub-task level skill labeling*.

2.2 Technical Skills Perception at Different Playback Speeds

Technical skill perception was measured by surveying non-expert crowds to give each video performance a GEARS score by rating each of the 5 subdomains shown in Table 1. Forty

Table 1: Likert-scale technical skill perception questionnaire, with five of the six domains of the GEARS assessment tool, excluding autonomy.

Score	**Depth Perception**
(1)	Constantly overshoots target, wide swings, slow to correct
(2)	
(3)	Some overshooting or missing of target, but quick to correct
(4)	
(5)	Accurately directs instruments in the correct plane to target

Score	**Bimanual Dexterity**
(1)	Uses only one hand, ignores non-dominant hand, poor coordination
(2)	
(3)	Uses both hands, but does not optimize interaction between hands
(4)	
(5)	Expertly uses both hands in a complementary way to provide optimal exposure

Score	**Efficiency**
(1)	Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress
(2)	
(3)	Slow, but planned movements are reasonably organized
(4)	
(5)	Confident, efficient and safe conduct, maintains focus on task, fluid progression

Score	**Force Sensitivity**
(1)	Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage
(2)	
(3)	Handles tissue reasonably well, minor trauma to adjacent tissue, rare suture breakage
(4)	
(5)	Applies appropriate tension, negligible injury to adjacent structures, no suture breakage

Score	**Robotic Control**
(1)	Consistently does not optimize view, hand position, or repeated collisions even with guidance
(2)	
(3)	View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant.
(4)	
(5)	Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant.

"turkers" were recruited independently for each video, in which each video at each play- back speed was independently submitted to the website in order to avoid a grouping bias. Videos were altered to play at

0.4x, 1.2x, 2.0x, 2.8x, and 3.6x, edited using FFmpeg [16], in which frames were either taken out or added in order to create the resulting playback speed.

The score from each of five subdomains (Depth Perception, Bimanual Dexterity, Efficiency, Force Sensitivity, and Robotic Control) were summed to create a cumulative score for each performance in the range of 5-25. Crowds were not specifically told whether the video was manipulated to play at a different speed, as this was not seen to make a difference in ratings for previous studies.

A linear mixed effects model was used to analyze the significance of the various speeds to the evaluations received. The significance of the time spent on reviewing a video and the labels given to the surgeons were also examined. It was hypothesized that each Mechanical Turker on the site should be assumed to have a different slope, to match the difference with which they relatively evaluate different videos. The evaluator ID given to each rater was assigned a random effect. In addition, fixed effects for both the speed at which the video was played and the amount of time spent on the video by the evaluator were controlled for in the model. The mixed model was compared with a null model, which did not include the fixed effects, using ANOVA hypothesis testing and comparing the relative information criteria and correlation between the scores and the various parameters. All data aggregation and statistics were calculated in Python 3.6 [17] and R [18].

2.3 Sub-Task Level Skill Labeling

To learn if the first minute of surgical activity displayed the same level of technical skill as the entire 15 minute video and should have the same label, the first minute of each video at 1x (normal playback speed) was submitted to crowds for review. The mean of these new scores were then calculated, and the 50th percentile was used as the cut-off point to signify in this dataset of 56 videos, whether a surgeon would be labeled as a proficient or expert level surgeon. The differences in these labels were then analyzed further and visualized, to find if the previously obtained scores for the entire 15 minute video possessed significantly different skill levels in the first minute, such that the label of the corresponding surgeon was different than what we had originally obtained.

3 Results

Table 2: Results from the linear mixed effects model testing speed and time spent on evaluation

Fixed Effects	**Estimate**	**Standard Error**	**df**	**t value**	**Pr (> | t |)**
Initial linear mixed effects model (BIC = 52465.36)					
Speed	0.046	0.013	13360	3.47	5.31e-04
Elapsed Time	-0.0023	3.47e-4	13720	-6.57	5.18e-11
Final linear mixed effects model (BIC = 52410.60)					

Table 3: The types of robotic surgery performed and demographics of the surgeons from the 56 videos used.

Surgery Type	N
Prostatectomy	17
Hysterectomy	11
Nephrectomy	5
Cystectomy	4
Other	19

Demographic	N
Male	17
Female	11
Mean Age	46.70
Mean Yrs. Experience	12.85

3.1 Data Demographics

Table 3 summarizes the types of robotic surgery and the demographic data from the $N = 56$ videos (and surgeons) used in this work.

3.2 Technical Skill Perception

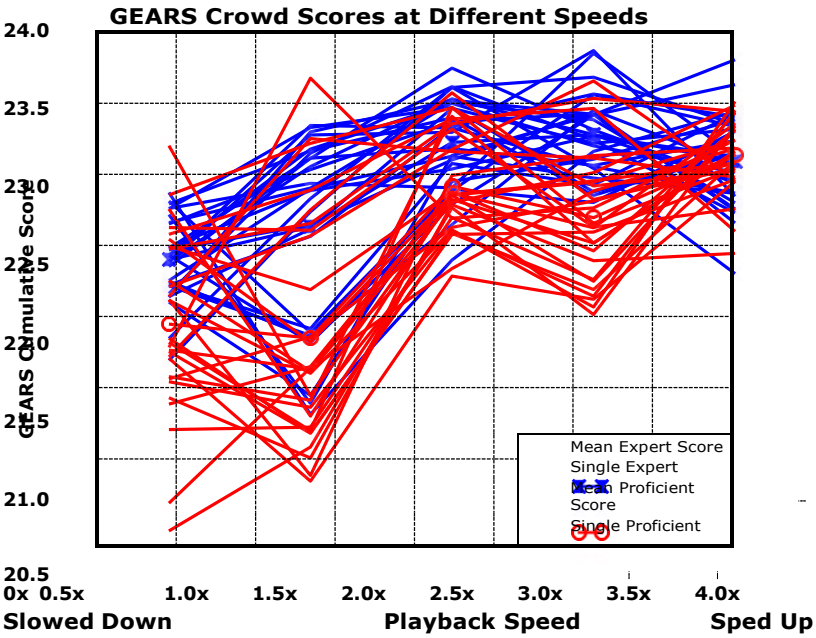
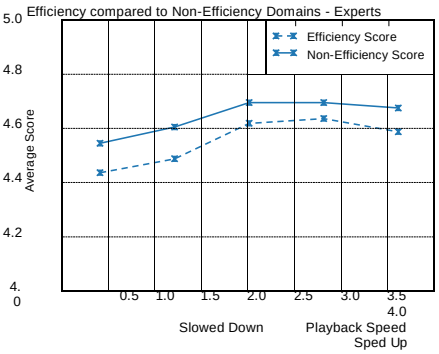


Fig. 3: All mean crowd evaluations (bold lines) from each proficient and expert surgeon at various video playback speeds. Each single surgeon's video (semi-transparent colors) indicates ratings from N = 40 turkers.


Efficiency compared to Non-Efficiency Domains - Experts
Efficiency Score
Non-Efficiency Score
Average Score
5.0
4.8
4.6
4.4
4.2
4.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Slowed Down
Playback Speed
Sped Up

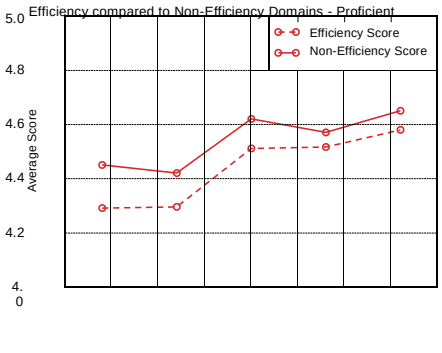

Efficiency compared to Non-Efficiency Domains - Proficient
Efficiency Score
Non-Efficiency Score
Average Score
5.0
4.8
4.6
4.4
4.2
4.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Slowed Down
Playback Speed
Sped Up

Fig. 4: The efficiency subdomain as compared to the mean of the other GEARS subdomains, for expert and proficient surgeons.

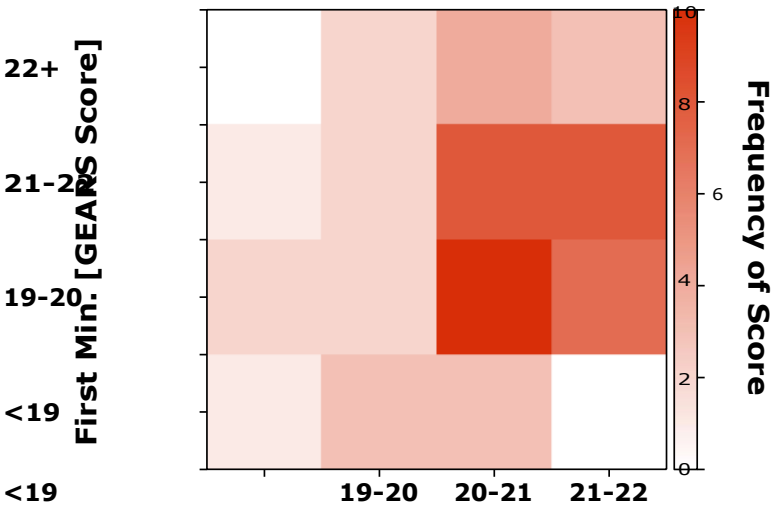
The mean of the GEARS score for each group and video are shown in Fig. 3. For expert videos, each increase in the playback speed by 0.8x was associated with, on average, a 0.16-point increase in the GEARS score (95% CI: 0.10-0.22 point increase; $p < 0.05$). On average these scores appear to increase within a sublevel of the playback speeds around 0.4x to 2.0x, and then level out at all remaining playback speeds. For proficient surgeon videos, each increase in the playback speed by 0.8x was associated with, on average, a 0.27-point increase in the GEARS score (95% CI: 0.19-0.35 point increase; $p < 0.05$). Thus, while both experts and proficient surgeons experienced increased perceived technical skill as the playback speed was increased, the mean score obtained at the fastest video playback speed reached an almost equal skill level for both proficient and expert level performances. Fig. 4 shows the increase in the efficiency subdomain as well as the mean of the other four domains to visualize whether efficiency (seen as the most related to speed) is the only increasing domain. As shown, no major difference is apparent between the two types of domains.

3.3 Sub-Task Level Labeling

Figure 5 illustrates the difference in GEARS scores by non-expert crowd workers, when viewing the first minute of the video compared to the entire video. As shown, there is a noticeable difference in the scores given at these two levels. Most surgeons, when viewed for the entire 15 minutes, are rated slightly higher than when only the first minute of surgical activity is performed. Viewing just the label given to the performance (proficient or expert), by analyzing whether the video was above or below the median score in the group of 56 videos, a total of 11 previously labeled proficient level surgeons, and 11 previously labeled expert surgeons switched the label they were originally given, when only the first minute of surgical activity was evaluated.

4 Conclusion

The results from the technical skill perception study give support to our initial hypothesis that increasing the video playback speed would increase the ratings of experienced surgeons. Now this evidence extends to real surgeries using robotic surgical performances.



Entire 15 Min. Video [GEARS Score]

Fig. 5: GEARS scores given to the entire 15-minute video of a performance, compared to only the first minute of the same video.

Surprisingly, however, we discovered that for sub-groups of expert level surgeons (proficient and expert) the increase in score happens at a quicker rate for proficient surgeons than for expert level surgeons. Increasing the playback speed of a slightly less than expert- level surgeon tends to “wash out” the minor mistakes they have made in the performance, effectively making the two groups appear more equally skilled at higher playback speeds. Additionally from analyzing the ‘Efficiency’ subdomain from the GEARS assessment, it appears that - surprisingly - crowds are also biased to give higher ratings in domains which are not associated with speed.

The results from the sub-task level skill labeling experiment show us that it may be necessary to have video evaluated in smaller duration segments, due to the notable disagreement with the skill level assigned to the entire video vs. just using the first minute. This lends support to the notion that a surgeon’s technical skill may fluctuate in time (on the order of minutes) throughout a surgical procedure. This warrants further study of non-constant skill mapped to a single summative rating, as this would induce substantial label noise of computational skill evaluation using machine learning.

We conclude that increasing the video playback speed of performances of practicing surgeons in typical robotic surgeries results in increased scores as reported by non-expert crowds. This effect is surprisingly uniform across GEARS subdomains, even those which should be unaffected by speed. We further conclude that more studies should be done to investigate variance in time of sub-task level videos of surgical procedures, as the technical skill of a surgeon may fluctuate on the order of minutes or less during the procedure.Compliance with Ethical Standards

Conflicts of Interest The authors declare that they have no conflict of interest.

Ethical Standard All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research commit- tee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Funding This work was supported, in part, by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-15-2-0030, the National Science Foundation CAREER grant under Award No. 1847610, as well as the National Institutes of Health’s National Center for Advancing Translational Sciences, grant UL1TR002494. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense, the National Science Foundation, or the National Institutes of Health’s National Center for Advancing Translational Sciences.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. Kowalewski T, Comstock B, Sweet R, Schaffhausen C, Menhadji A, Averch T, Box G, Brand T, Fer- randino M, Kaouk J, Knudsen B, Landman J, Lee B, Schwartz BF, McDougall E, and Lendvay TS (2015) Crowd-Sourced Assessment of Technical Skills for Validation of Basic Laparoscopic Urologic Skills (BLUS) Tasks. The Journal of Urology 195(6):1859–1865.
2. Kelly JD, Petersen A, Lendvay TS, and Kowalewski TM (2020) The Effect of Video Playback Speed on Surgeon Technical Skill Perception. International Journal of Computer Assisted Radiology and Surgery
3. Makary M and Daniel M (2016) Medical error - the third leading cause of death in the US. The BMJ.
4. Birkmeyer JD, Finks JF, O’Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, and Birkmeyer NJ (2013) Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine 369(15):1434–1442.
5. Dockter R, Lendvay TS, Sweet RM, and Kowalewski TM (2017) The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. The International Journal of Computer Assisted Radiology and Surgery 12:1151–1159.
6. Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B, Kuksenok K, Aragon C, Holst D, and Lendvay T (2013) Crowd-Sourced

Seattle Children's — Internal

Assessment of Technical Skills: a Novel Method to Evaluate Surgical Performance. Journal of Surgical Research 187(1):65–71.
7. Doh FC, Goldfarb DW, Sander JC, Miles BJ, and Dunkin BJ (2012) Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. The Journal of Urology 187(1):247–252.
8. Kowalewski TM, White LW, Lendvay TS, Jiang IS, Sweet RS, Wright A, Hannaford B, and Sinanan MN (2014) Beyond task time: automated measurements augments fundamentals of laparoscopic skills methodology. Journal of Surgical Research 192(2): 329–338.
9. Hung A, Chen J, Che Z, Nilanon T, Jarc A, Titus M, Oh PJ, Gill IS, and Liu Y (2018) Utilizing Ma- chine Learning and Automated Performance Metrics to Evaluate Robot-Assisted Radical Prostatectomy Performance and Predict Outcomes. Journal of Endourology 32(5).
10. Jokisch D, Troje NF (2003) Biological motion as a cue for the perception of size. Journal of Vision 3(1): 252–264.
11. Veto P, Einhauser W, and Troje NF (2017) Biological motion distorts size perception. Scientific Reports 7(42576):1–6.
12. Seete RM, Beach R, Sainfort F, Gupta P, Reihsen T, Poniatowski LH, and McDougall EM (2012) In- troduction and validation of the American Urological Association Basic Laparoscopic Urology Surgery skills curriculum. Journal of Endourology, 26: 190.
13. Kowalewski TM, Seet R, Lendvay TS, Menhadji A, Averch T, Box G, Brand T, Ferrandino M, Kaouk J, Knudsen B, Landman J, Lee B, Schwartz BF, and McDougall E (2016) Validation of the AUA BLUS Tasks. Journal of Urology 195: 998.
14. Satava RM, Gallagher AG, and Pellegrini CA (2003) Surgical competence and surgical proficiency: definitions, taxonomy, and metrics. Journal of the American College of Surgeons 196(6): 933-7.
15. French A, Lendvay TS, Sweet RM, and Kowalewski TM (2017) Predicting surgical skill from the first n seconds of a task: Value over task time using the isogony principle. The International Journal of Com- puter Assisted Radiology and Surgery 12(7): 1161–1170.
16. FFmpeg Developers (2016) ffmpeg tool (Version 4.1.3) [Software]. Available form http://ffmpeg.org/.
17. Python Software Foundation. Python Language Reference, version 3.6. Available at http://www.python.org
18. R Core Team (2018) R: A Language and Environment for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Appendix G

Int J CARS manuscript No.
(will be inserted by the editor)

Temporal Variability of Surgical Technical Skill Perception in Real Robotic Surgery

Jason D. Kelly, Michael Nash M.S., Nicholas Heller, Thomas S. Lendvay M.D., Timothy M. Kowalewski Ph.D.

Received: date / Accepted: date

Abstract *Purpose:* Summary score metrics, either from crowds of non-experts, faculty surgeons, or automated performance metrics (APMs), have been trusted as the prevailing method of reporting surgeon technical skill. The aim of this paper is to learn whether there exist significant fluctuations in the technical skill assessments of a surgeon throughout long durations of surgical footage.
Methods: A set of 12 videos of common robotic surgery cases from human patient robotic surgeries were used to evaluate the perceived technical skill at each individual minute of the surgical videos, which were originally 12-15 minutes in length. A linear mixed effects model for each video was used to compare the ratings of each minute to the overall mean for the entire video in order to learn whether a change in scores over time is more significant than typical intrarater variation. *Results:* Modeling the change over time of the Global Evaluative Assessment of Robotic Skills (GEARS) scores significantly contributed to the prediction model. This gives evidence that significant changes in the technical skill over time are noticeable in the model.
Conclusion: The findings from this research raise questions about the optimal duration of footage needed to be evaluated to arrive at an accurate rating of surgical technical skill for longer procedures. This may imply non-negligible label noise for supervised machine learn- ing approaches. In the future, it may be necessary to report a surgeon's skill variability in addition to their mean score to have proper knowledge of a surgeon's overall skill level.

Seattle Children's — Internal

Keywords Crowd Sourcing · Surgical Technical Skill · Video Segmentation · Bias

Jason Kelly
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN E-mail: kell1917@umn.edu

M. Nash
Department of Biostatistics, University of Washington, Seattle, WA

N. Heller
Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN

T. Lendvay
Department of Urology, University of Washington, Seattle, WA

T. Kowalewski
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN

1 Introduction

Methods for assessing the technical skills of surgeons is paramount to ensuring to the public that surgeons are safe and effective. For years summary scores or metrics have been used as the main method to report surgical skill. The most popular of these has been to use a likert-scale scoring metric, from either non-expert crowds or faculty surgeons. Past research has shown that crowds of non-experts concord with surgeon raters in evaluating technical skill [1]. Automated performance metrics have also been used, in which surgical events or streaming kinematic data have been used for computation of various metrics, suggesting superior objectivity [2]. As it is known that surgical skill is related to patient outcomes, and medical errors are the third leading cause of death in the United States, for which surgical errors contribute a large part, it remains important to accurately assess and report surgeon skill [3], [4]. Much progress has been made by using statistical and machine learning models in the past with the goal of classifying surgeons into skill levels of 'novice' and 'expert' [5], but it remains a difficult computational task. For the largest statistical power this usually leaves the most reliable approach being to obtain either surgeon or crowd evaluations from video [6].

One of the most popular robotic surgical skill assessment metrics is the Global Evaluative Assessment of Robotic Skills (GEARS), which is the most common objective assessment tool for robotic surgery [7]. The subdomains in this metric include: bimanual dexterity, efficiency, depth perception, force sensitivity and robotic control. Scores of 1-5 for each of these subdomains are totaled, for a cumulative score of 5-25, with higher scores belonging to more skilled surgeons. There is an additional domain in GEARS named Autonomy which is not used here. Autonomy is typically used when evaluating a surgeon's ability to work independently, which can't be accurately assessed through video alone.

At least two confounding factors arise: (1) limitation of human attention span of the raters which inhibits reliability of their ratings and (2) the fluctuations of "true technical skill", naturally exhibited by a surgeon through time, when videos of a longer duration are used. It is unclear whether ratings remain reliable as durations scale up. The former is akin to *measurement noise* and is an attribute of the method to measuring skill - be it human ratings or computed computationally. The latter is akin to *process noise* and is an attribute of the surgeon's activities and tool-tissue environment interactions. Studies of human attention span have found that people viewing video lectures on average experienced significant decreases in attention and even stopped viewing after only about 6 minutes. One specific study found that including interactivity elements to video lectures led to an increase in watch time by at least 20% [8]. Given that evaluating surgeons requires an even higher degree of focus, using videos which are longer than 5 minutes may be detrimental to receiving accurate evaluations. Additionally, when making judgments, crowds can be susceptible to a contrast effect bias in which wide ranges in performance can lead to wildly different evaluations [9]. This could mean a performance of a novice who performed unusually well in the last segment of a video may receive unreliably high ratings. The natural fluctuations in technical skill remain largely unexplored, though hinted at in [10], [11].

Seattle Children's — Internal

We hypothesize that statistically significant degrees of temporal fluctuations of perceived skill exist between smaller segments of surgical performances, which differ from previously obtained summary scores of the longer duration videos.

Table 1: Likert-scale technical skill perception questionnaire, from the five domains of the GEARS assessment metric.

Score	**Depth Perception**
(1)	Constantly overshoots target, wide swings, slow to correct
(2)	
(3)	Some overshooting or missing of target, but quick to correct
(4)	
(5)	Accurately directs instruments in the correct plane to target
Score	**Bimanual Dexterity**
(1)	Uses only one hand, ignores non-dominant hand, poor coordination
(2)	
(3)	Uses both hands, but does not optimize interaction between hands
(4)	
(5)	Expertly uses both hands in a complementary way to provide optimal exposure
Score	**Efficiency**
(1)	Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress
(2)	
(3)	Slow, but planned movements are reasonably organized
(4)	
(5)	Confident, efficient and safe conduct, maintains focus on task, fluid progression
Score	**Force Sensitivity**
(1)	Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage
(2)	
(3)	Handles tissue reasonably well, minor trauma to adjacent tissue, rare suture breakage
(4)	
(5)	Applies appropriate tension, negligible injury to adjacent structures, no suture breakage
Score	**Robotic Control**
(1)	Consistently does not optimize view, hand position, or repeated collisions even with guidance
(2)	
(3)	View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant.
(4)	
(5)	Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant.

2 Methods

2.1 Dataset

This study utilized 12 videos from the Robotic Surgery Readiness (RSR) study. The goal of that study was to determine whether pre-operative warm-up on virtual reality tasks had a

Group	Sub-Group	Videos (N)	Prostatectomy	Hysterectomy	Other
Expert	Warmup	3	1	1	1
	Control	3	1	0	2
Proficient	Warmup	3	2	0	1
	Control	3	1	1	1

Seattle Children's — Internal

Table 2: Summary information for 12 videos used from the RSR study.

measurable improvement in surgical technical skills among practicing surgeons (no novice trainees) after typical periods of surgical inactivity, in a controlled-randomized trial. This dataset contains 343 videos of surgical procedures consisting of live robotic surgeries, which were performed by forty attending surgeons and trainees in urology, gynecology, and general surgery at the University of Washington Medical Center and the Madigan Army Medical Center. These robotic surgeries were performed using da Vinci surgical robots, created by Intuitive Surgical (Sunnyvale, CA). Each video was manually edited to include roughly the first 15 minutes of surgical activity performed by the criterion surgeon. This beginning portion of the surgery contained similar actions among the different surgery specialties, such as suturing and cutting actions. A GEARS score for each 12-15-minute performance was previously obtained from CSATS-Inc., using crowd evaluation.

The range of scores in all RSR videos was fairly small, with most lying between 20-22 out of 25 (only 15% of the full range possible). This was the case as all surgeons assessed were practicing faculty and no trainees were involved in these videos. To obtain the largest possible range of skill from this dataset, 6 performances from the top quintile of scored performances and 6 from the bottom quintile of performances were used and given labels of ‘expert’ and ‘proficient’, respectively, keeping in mind that all surgeons in this dataset are highly-skilled, or they wouldn’t be allowed to operate in live robotic surgery procedures [12]. Additionally, half of each group of videos were performances in which the surgeon participated in a simulated surgery warm up procedure beforehand, whereas the other half started surgery without warming up. This stratified formatting is illustrated in Table 2.

2.2 Crowd Evaluation

Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each non-expert crowd worker was paid $0.50 to watch and evaluate each of a series of one-minute videos. A web domain was created for which Turkers would be redirected, where they submitted a consent form and were asked GEARS questions about videos. Each 12-15-minute video performance was segmented into smaller videos having a duration of one minute, using FFMPEG, an open source video editing tool [13]. The order of these videos were randomized so that crowds were viewing a non-chronological ordering. They were asked to provide a GEARS score for each minute-long video and proceeded to evaluate all remaining videos for the same performance. One-minute segments were used as this pro- vided a higher level of granularity to see differences in ratings throughout the performance. Smaller segments of videos, such as 30 seconds, appeared to be too short and sometimes videos lacked meaningful surgical activity in these smaller segments. Forty raters were recruited to rate each performance.

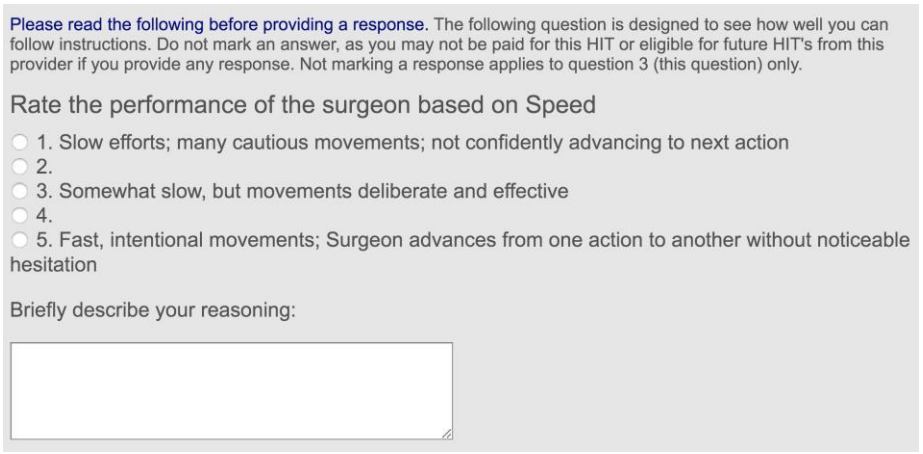


Fig. 1: The attention question inserted to the GEARS questionnaire, for quality assurance purposes. Note that skimming the instructions likely results in incorrect answers.

Seattle Children's — Internal

A few extra elements were added to the user interface in an effort to increase interactivity and obtain better data quality. Once a crowd worker gave consent to participating in the study, they were shown a video with two performances, side by side, of a novice and expert performer of a laparoscopic surgical training task. The crowds were asked to evaluate which surgeon was better, and to what degree. After evaluating this task, the series of robotic surgery procedures began. After each GEARS likert-scale question prompt, a text box was inserted, asking the raters to provide reasoning for their decision in each subdomain. These elements, such as the rater training portion of the study, have shown us in the past that we obtain more reliable ratings if these are included, as this might cause crowd workers to feel their ratings are more valued than when these elements are excluded.

Finally, an additional question, which is not part of the standard GEARS questionnaire was inserted, asking for a rating of the 'speed' of the surgeon, shown in Fig. 1. Text in the prompt stated that the worker should not answer this question. This 'attention question' served as a mechanism to query which turkers were focusing on the questionnaire vs. those quickly finishing to be paid. All ratings which included answers to this 'attention question' were removed from analysis.

2.3 Statistical Analysis

A linear mixed effects model for each video was used to analyze the outcome of GEARS scores assigned by each rater at each timepoint. A linear mixed effects model was chosen as these are often used for biological testing with nested groups of responses. A random intercept was estimated for each individual rater to model rater-by-rater variability in scores. Fixed effects were also estimated for each minute of video in order to quantify differences in surgical skill over time. With each model, we tested the hypothesis that the mean GEARS score was the same at each timepoint to determine if there were significant changes in skill over time for each surgeon. All statistical work was accomplished using R 3.6.2 [14], with data manipulation and visualization computed in Python 3.6 [15]. Results

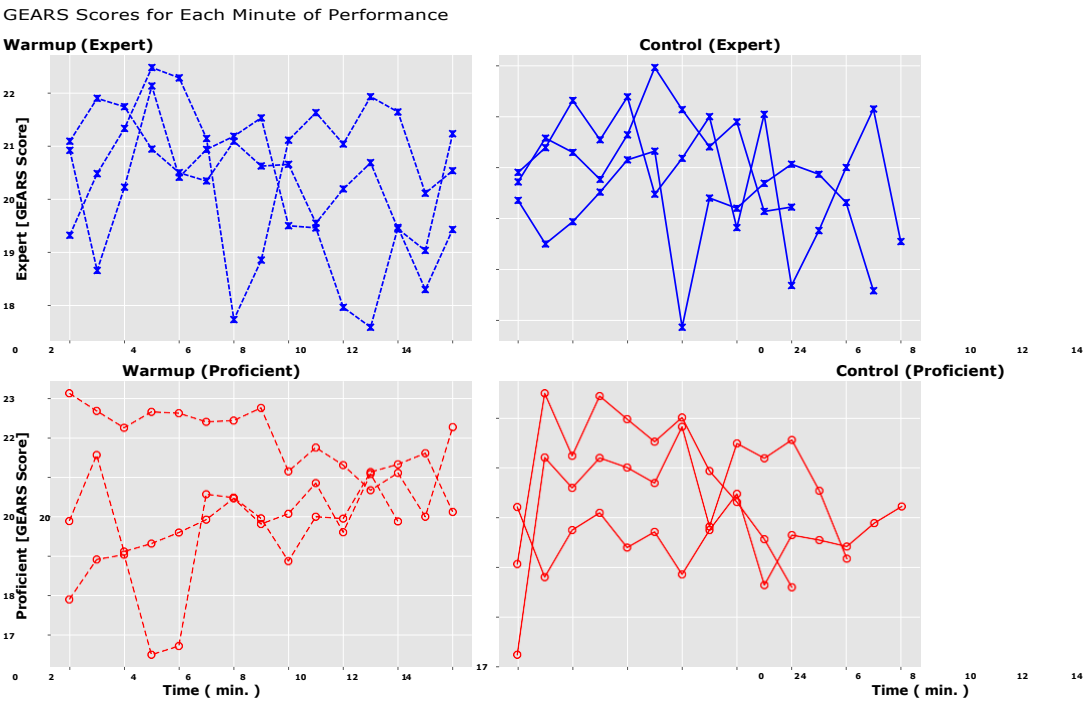


Fig. 2: All mean crowd scores (N=40 per dot) from each proficient (red) and expert (blue) surgeon at each minute of the performance. Surgeons previously randomized to the control (solid lines) group normally started surgery without any intervention or pre-operative warm-up used. Warm-up (dashed lines) group surgeons reviewed a virtual reality warm-up module prior to surgery. (The warm-up hypothesis from the original randomized study is not being tested or evaluated in this research, only temporal variation in ratings.)

The mean of the GEARS score for each expert and proficient surgeon are shown in Fig. 2. No identifiable relationship or trend in any of the 4 subplots is immediately apparent. However, significant differences in the mean score over time were detected for ten of the twelve videos ($p = 0.001891$ to $2.531e\ 14$) using linear mixed effects models. For two videos, the effect of time was not significant (proficient control video 1, $p = 0.61794$, and proficient warm-up video 2, $p = 0.2618$).

A bar plot comparing the previously obtained summary score for each surgeon, grouped with the mean of each 1-minute segment, is displayed in Fig. 3. Several of the previously obtained summary scores are outside the confidence interval band from the mean of the one-minute scores. Overall, we saw each full duration video's 1-minute GEARS mean scores vary by 1.02 ± 0.28 points.

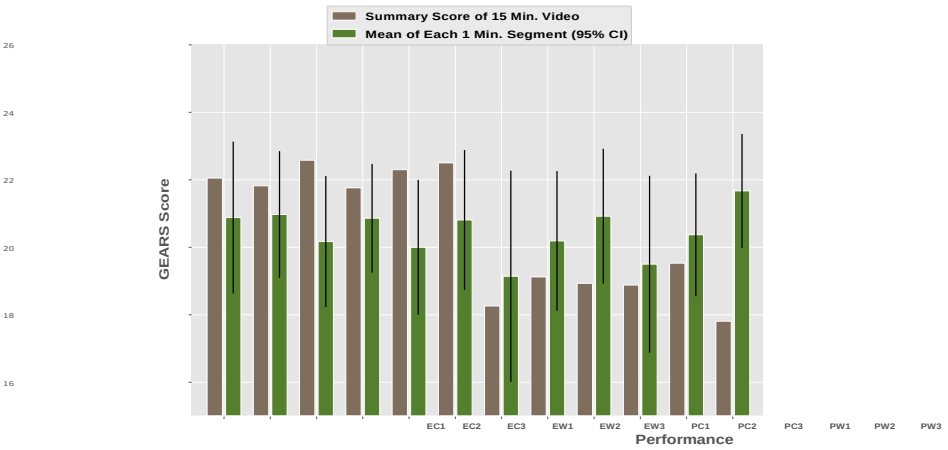


Fig. 3: Comparison of previously obtained scores for entire 15 minute video to the mean score for all 15 1-minute segments (95% CI) in the same video. The confidence intervals were not available from the vendor used to obtain scores for 15-minute segments. (PW = Proficient/Warmup; PC = Proficient/Control; EW = Expert/Warmup; EC = Expert/Control)

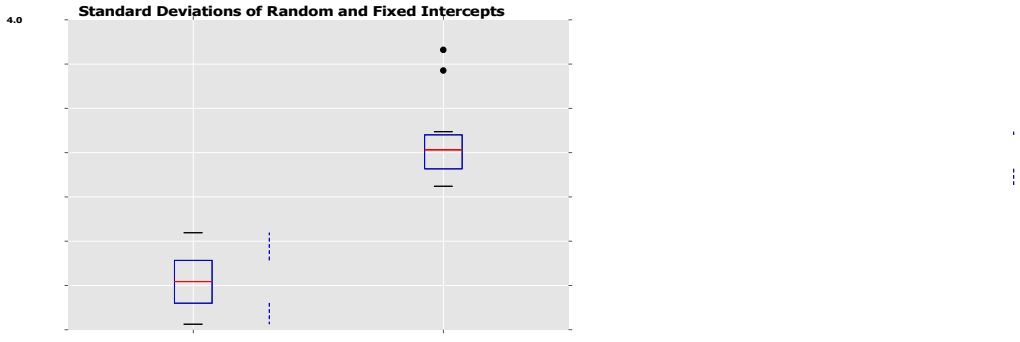




Fig. 4: Standard deviation from the LME model for random effects from each individual crowd rater, compared with the fixed effect of time. The rater effect is clearly larger, although time is still significant in predicting outcomes.

3 Conclusion

The results from the video segmentation study support our initial hypothesis that perceived surgeon skill may fluctuate throughout surgical procedures. If we consider the limitations of human attention span and time-varying fluctuations in skill for a 15-minute video as contributing factors, it appears that human attention span limitations have a larger effect than natural variation of "true" technical skill in time (i.e. measurement noise may be 2- 3x larger than process noise). We found no compelling evidence that the underlying human

skill (process) is constant for a 15-minute duration in typical surgery. Perhaps crowds may be biased into giving scores that more closely reflect events they remember as being particularly good or bad. This research may give credit to alternative surgical skill reporting methods, as opposed to giving static scores and labels. One alternative to this could be providing the mean of shorter segments of videos, as well as the amount of deviation in score throughout the performance, as a confidence interval. This would convey both the typical skill of a surgeon in addition to how consistent their skill is through time.

This study includes a few limitations. There was no semantic segmentation of these tasks, only temporal "chunking" into one-minute segments. Although videos of real surgeries were used, the data used was fairly sparse, with only 12 videos of 15-minute surgeries. It is also not clear if the one-minute duration used for segmentation is optimal. Shorter or longer segments of surgery may be more beneficial for accurate skill evaluation. In this work, the one-minute ratings were obtained 3 months after the full duration videos were rated. This may raise concerns about reliability (and comparability of data) of crowd ratings from studies separated by large time intervals.

Using these alternative reporting methods could aid supervised machine learning models in the future by reducing potential label noise. Static scores may lead to noisily-labeled data sets, producing poorly trained classification models. Further study should consider optimal duration for human raters and a more rigorous analysis of the natural fluctuations of human technical skill.

Compliance with Ethical Standards

Conflicts of Interest The authors declare that they have no conflict of interest.

Ethical Standard All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research commit- tee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Funding This work was supported, in part, by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-15-2-0030, the National Science Foundation CAREER grant under Award No. 1847610. Opinions, interpretations, conclu-ions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense, the National Science Foundation,

[Type here]

or the National Institutes of Health's National Center for Advancing Translational Sciences.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. Kowalewski T, Comstock B, Sweet R, Schaffhausen C, Menhadji A, Averch T, Box G, Brand T, Ferrandino M, Kaouk J, Knudsen B, Landman J, Lee B, Schwartz BF, McDougall E, and Lendvay TS (2015) Crowd-Sourced Assessment of Technical Skills for Validation of Basic Laparoscopic Urologic Skills (BLUS) Tasks. The Journal of Urology 195(6):1859–1865.
2. Hung A, Chen J, Che Z, Nilanon T, Jarc A, Titus M, Oh PJ, Gill IS, and Liu Y (2018) Utilizing Machine Learning and Automated Performance Metrics to Evaluate Robot-Assisted Radical Prostatectomy Performance and Predict Outcomes. Journal of Endourology 32(5).
3. Makary M and Daniel M (2016) Medical error - the third leading cause of death in the US. The BMJ.
4. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, and Birkmeyer NJ (2013) Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine 369(15):1434–1442.
5. Kelly JD, Petersen A, Lendvay TS, and Kowalewski TM (2020) The Effect of Video Playback Speed on Surgeon Technical Skill Perception. International Journal of Computer Assisted Radiology and Surgery.
6. Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B, Kuksenok K, Aragon C, Holst D, and Lendvay T (2013) Crowd-Sourced Assessment of Technical Skills: a Novel Method to Evaluate Surgical Performance. Journal of Surgical Research 187(1):65–71.
7. Doh FC, Goldfarb DW, Sander JC, Miles BJ, and Dunkin BJ (2012) Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. The Journal of Urology 187(1):247–252.
8. Geri N, Winer A, Zaks B (2017) Challenging the six-minute myth of online video lectures: Can inter- activity expand the attention of learners? Online Journal of Applied Knowledge Management 5(1):101– 111.
9. Schmitt M, Bulterman DCA, Cesar PS (2018) The contrast effect: QoE of mixed-video qualities at the same time. Quality and User Experience 3(7).
10. French A, Lendvay TS, Sweet RM, and Kowalewski TM (2017) Predicting surgical skill from the first n seconds of a task: Value over task time using the isogony principle. The International Journal of Computer Assisted Radiology and Surgery 12(7): 1161–1170.
11. Kelly JD, Heller N, Petersen A, Lendvay TS, and Kowalewski TM (2020) The Effect of Video Play- back Speed on Perception of Technical Skill in Robotic Surgery. The International Journal of Computer Assisted Radiology and Surgery [Submitted].
12. Dockter R, Lendvay TS, Sweet RM, and Kowalewski TM (2017) The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. The International Journal of Computer-Assisted Radiology and Surgery 12:1151–1159.
13. FFmpeg Developers (2016) ffmpeg tool (Version 4.1.3) [Software]. Available form http://ffmpeg.org/.
14. R Core Team (2018) R: A Language and Environment for Statistical Computing, Vienna, Austria. http://www.R-project.org/
15. Python Software Foundation. Python Language Reference, version 3.6. Available at http://www.python.org

19. cDougall E (2016) Validation of the AUA BLUS Tasks. Journal of Urology 195: 998.
20. Satava RM, Gallagher AG, and Pellegrini CA (2003) Surgical competence and surgical proficiency: definitions, taxonomy, and metrics. Journal of the American College of Surgeons 196(6): 933-7.
21. French A, Lendvay TS, Sweet RM, and Kowalewski TM (2017) Predicting surgical skill from the first n seconds of a task: Value over task time using the isogony principle. The International Journal of Com- puter Assisted Radiology and Surgery 12(7): 1161–1170.
22. FFmpeg Developers (2016) ffmpeg tool (Version 4.1.3) [Software]. Available form http://ffmpeg.org/.
23. Python Software Foundation. Python Language Reference, version 3.6. Available at http://www.python.org
24. R Core Team (2018) R: A Language and Environment for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Appendix H

Robot-Assisted Surgery Readiness: Virtual Reality Warm-Up Prior to Robot-Assisted Surgery

Jason D. Kelly[1] M.S., Timothy M. Kowalewski[1] Ph.D., Tim Brand[2] M.D., Anna French[1] Ph.D., Michael Nash[2] M.S., Lois Meryman[2] B.S., Nicholas Heller[1] B.S., Nancy Organ[2] M.S., Evalyn George[3] B.S., Roger Smith[4] M.D., Mathew D Sorensen[2] M.D., Bryan Comstock[2] M.S., Thomas S. Lendvay[2] M.D.

AUTHOR INFORMATION

1. University of Minnesota, Minneapolis, MN, USA
2. University of Washington, Seattle, WA, USA
3. Madigan Army Medical Center, Henry M. Jackson Foundation, Tacoma, WA, USA
4. Florida Hospital Nicholson Center, Orlando, FL, USA

Corresponding Author

* Jason D. Kelly

Present Address

1501 Portland Ave. Apt. 407

Minneapolis, MN, 55404

Email: kell1917@umn.edu

Manuscript Word Count: 2932

Key Points

Question: What is the optimal virtual reality warm-up curriculum to elevate technical skill, and does this curriculum elevate technical skills proficiency in robot-assisted surgery?

Findings: In this randomized clinical study that included 41 surgeons it was found that

performing a brief virtual reality warm-up module prior to performing robot-assisted surgery does not significantly improve the technical skill of robotic surgeons, as measured by video review and objective metrics.

Meaning: Virtual reality warm-up prior to robot-assisted surgery may not improve the technical skill of practicing surgeons.

ABSTRACT:

Importance: *Intra-operative errors correlate with surgeon skill and skill declines with intervals of inactivity.*

Objective: *The goals of this research were to identify the optimal virtual reality (VR) warm-up curriculum to prime a surgeon's technical skill and validate benefit in the operating room.*

Design: *Surgeons were randomized to receive six trial sessions of a designated set of VR modules on the da Vinci Skills Simulator. After performing their curricula, warm-up effect was assessed based on performance on a criterion task. The optimal warm-up curriculum was chosen from the group with the best task time and video review-based technical skill. Robot-assisted surgery (RAS)-experienced surgeons were then recruited to either receive or not receive warm-up before surgery. Skill in the first 15 minutes of surgery was assessed byblinded surgeon and crowd-worker review as well as tool motion metrics.*

Setting: *Surgeons and trainees were recruited from urology, gynecology and general surgery from military, Veteran's Administration, public academic, and private medical centers.* .

Participants: Forty-one surgeons participated in the warm-up curriculum selection, while 34 practicing RAS-experienced surgeons participated in intra-operative warm-up validation. .

Intervention: The intervention was performing VR warm-up prior to human RAS.

Main Outcome and Measure: Warm-up effect was measured using objective performance metrics and video review using the Global Evaluative Assessment of Robotic Skills (GEARS) tool. Linear mixed effects models with a random intercept for each surgeon and nonparametric modified Friedman tests were used for analysis.

Results: A Running Suture task on the simulator was on average 31.3 seconds faster, and had the highest GEARS scores, thus was chosen as the optimal curriculum. Thereafter, surgeons completed 343 surgeries with corresponding video and tool motion data. No statistically significant differences in skill were observed with the warm-up intervention. .

Conclusion and Relevance: We conclude that a robotic VR warm-up prior to performing the early stages of surgery does not impact he technical skill of the surgeon.

Disclosure Information: Nothing to disclose.

This study was supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-15-2-0030. The views, opinions, and endorsement by the author(s) do not reflect those of the US Army or the Department of Defense.
The Seattle Children's Core for Biomedical Statistics is supported by the Center for Clinical and Translational Research at Seattle Children's Research Institute and grant UL1RR025014 from the NIH National Center for Research Resources.

Author Contributions

Study conception and design: Lendvay, Sorensen, Smith, Brand, and Kowalewski

Acquisition of data: Meryman

Analysis and interpretation of data: Kelly, French, Nash, Heller, Organ, George, Comstock, and Kowalewski

Drafting of manuscript: Kelly, Meryman, Heller, Nash, Lendvay, and Kowalewski

ABBREVIATIONS

GEARS = Global Evaluative Assessment of Robotic Skills
UWMC = University of Washington Medical Center
NAD = Normalized Angular Displacement
API = Application Programming Interface
MAMC = Madigan Army Medical Center
ROC = Rate of Orientation Change
RSR = Robotic Surgery Readiness
CUP = Camera Use Per Minute
UMN = University of Minnesota
RAS = Robot-Assisted Surgery
VA = Veterans Administration
WUD = Warm-Up Decrement
IRB = Internal Review Board
SAL = Spectral Arc Length
OR = Operating Room
MV = Mean Velocity
VR = Virtual Reality
TT = Task Time

INTRODUCTION

Medical errors account for 98,000 deaths each year, atop a higher number of patient complications.[1] An estimated one third of these errors are surgical and the average American can expect to have 7 surgeries in their lifetime.[2-4] Although surgical simulation's role in off-line teaching is accepted, [5] simulation may also have a role in identifying decrements in surgical performance and offer a system for mitigating skill decay.[6-9] Perez et al. concluded that skill may decay at different degrees depending on the type of skill, and that training to mitigate decay may require targeted curricula based on the specific decay signatures.[7] These methods were described for laparoscopic surgery skills, yet no warm-up has been studied for robot-assisted surgery(RAS) which has seen considerable adoption in the United States with the creation of the da Vinci surgical robot by Intuitive Surgical(Sunnyvale, CA).[8] .

While several other checklist-type preparation procedures exist in the operating room(OR), there is still no metric which measures how prepared a surgeon is following periods of inactivity. [9] In other professions, after participating in a warm-up, task performance can be elevated.

[10] Previous work additionally shows that a warm-up decrement (WUD) exists, in which the effects of warm-up begin decaying approximately 30 minutes after performing a warm-up task. [10-13] These hypotheses have already found four factors which most affect WUD: (1) intervals between practice, (2) complexity and type of task, (3) strategies employed to maintain skills, and (4) individual differences. [14,15] Virtual reality(VR) proficiency-based warm-up modules have been shown to most effectively combat skill decay. [16-20] This research group identified in a prior study that VR warm-up primes experienced robotic surgeons when looking at task time and errors, thus we sought to assess the warm-up benefit among practicing surgeons.[19

The da Vinci Skills Simulator (dVSS) replaces the physical robotic arms of a da Vinci robot with virtual representations of the tools to simulate a surgical environment (eFigure 1). [21] The simulator tasks exercise a surgeons ability to move objects, target camera position, dissect tissues, and suture.

eFigure 1 The da Vinci Skills Simulator allows surgeons to practice typical surgical tasks without the danger of performing in real surgery.



Measuring surgical technical skill in an objective manner is difficult, with the gold standard being video review by expert surgeons. Previous work has shown that video review of technical skill by non-expert crowds is accurate. When compared to expert surgeon review, crowds of non-experts were able to predict the pass/fail rate of surgeons with 100% accuracy in retrospective studies. [22,23] Tool motion data captured from the da Vinci has also been associated with surgical skill. Metrics such as Task Time TT), spectral arc length (SAL), camera use per minute (CUP), normalized angular displacement (NAD), rate of orientation change (ROC), and mean velocity (MV), have all been found to associate with surgical skill. [24-28
In this study we sought to find the optimal VR warm-up curriculum to properly prime

surgeons for RAS and validate the warm-up benefit in the OR by measuring objective performance metrics and obtaining video review of participants.

METHODS

AIM 1

Subject Recruitment

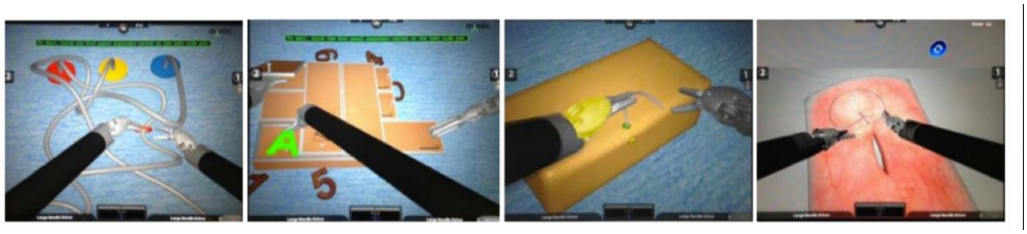
Surgeons and trainees were recruited after Institutional Review Board approval(#41730) from urology, gynecology and general surgery at four clinical sites including military, Veteran's Administration Hospital, public academic, and private medical centers. Each of the subjects were asked to complete a standard demographics questionnaire to learn their experience and any factors impacting responses to warm-up simulations. The information obtained is shown in eTable 1.

eTable I List of demographics information obtained from enrolled participants in the study.

Information
Unique Subject Code
Age
Gender
Years in training
Years since graduated training
Surgical specialty
Robotic surgery exp. (# cases as primary surgeon)
Laparoscopic surgery exp. (# cases as primary surgeon)
Handedness
Musical experience
Date of return since last deployment
Duration of deployment
Surgical role while deployed (if applicable)
Time since last performed robot-assisted surgery or procedure

dVSS Cirricula

Modules used for surgeon priming needed to be technically challenging to stimulate the surgeon in a way that most closely corresponds to movements present in a surgical environment. The criteria used to select modules were: (1) prior validation in the literature, and/or 2) involvement of multiple skills simultaneously, and/or 3) containing content that closely simulates actual surgery. Following these criteria, we chose to use the Ring and Rail 2, and Match Board 3 modules from the EndoWrist Manipulation 2 category. Ring and Rail 2" has previously been found to effectively discriminate surgical skills, and Match Board 3 has been found to be the most difficult of the dVSS modules among experienced surgeons. [17,29]. From the Needle Driving category, Suture Sponge 3 was chosen as it has been found to vary greatly between experience, intermediate, and novice performers.[29] "Running Suture 2" was the final module chosen, and used as the criterion module against which all groups of subjects were tested for proficiency, as it most closely represents the actions in a real surgical suturing procedure. Each of these four modules can be completed in approximately 2-3 minutes, making them short enough to be feasibly used as a component of a final curriculum.



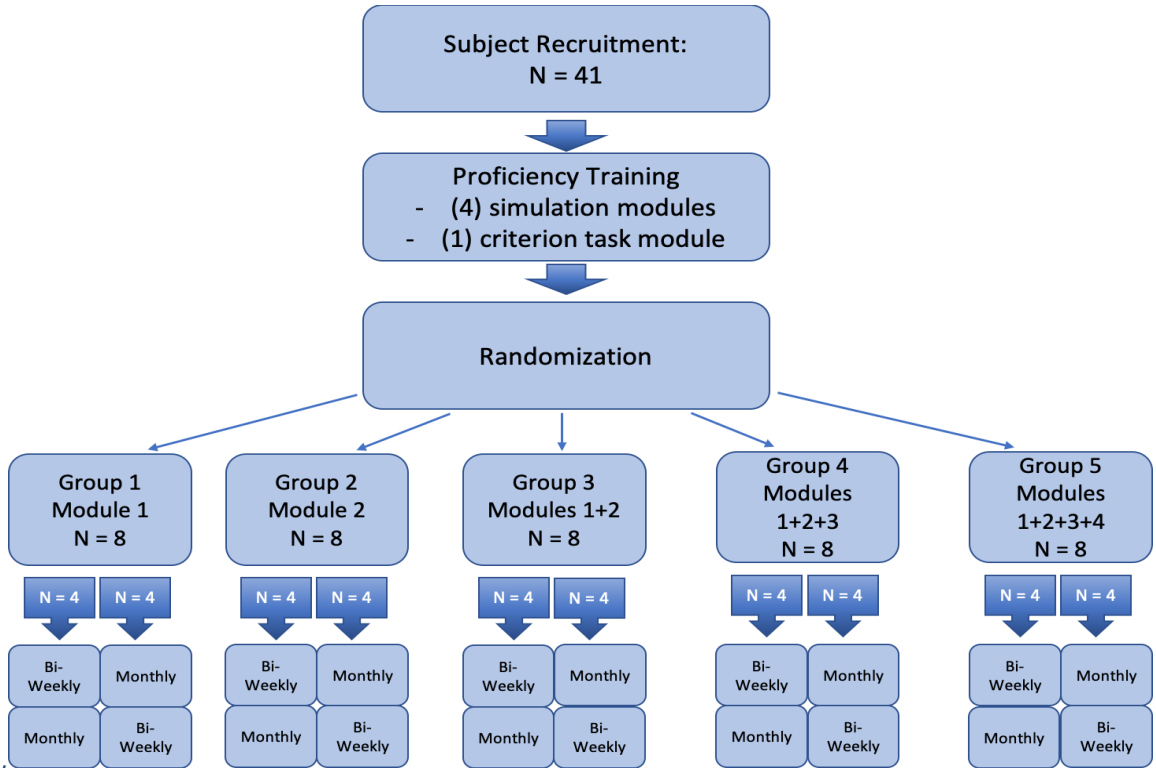
eFigure 2 Frames from each of the four warm-up modules used during Aim 1. From left to right: Ring and Rail, Match Board, Suture Sponge, and Running Suture.

Proficiency Testing

To normalize the skills of subjects, each underwent proficiency training before being randomized to one of 5 curricula groups (eFigure 2). Proficiency testing included completion of the Intuitive Surgical

da Vinci didactic web-based curriculum. To obtain benchmark performance goals for the proficiency training, two expert robot-assisted surgeons completed the four modules used until they made no errors. Subjects were deemed “proficient” once they met task time benchmarks for three consecutive trials.

eFigure 3 Randomization protocol for selection of warm-up module for surgeons



The four dVSS modules were split into five curricula groups with different combinations of modules to test which curriculum was most efficient for priming surgeons. The modules in each of these five groups are shown in Table 1. Within each group, half were randomized to perform the tasks monthly and half performed bi-weekly. After 8 weeks of performing at one frequency, the participants were then crossed-over to the opposing frequency group for a total of six trial sessions, to allow for certain baselines in statistical models to control for intervals of inactivity. After performing his/her respective curriculum, each participant performed the criterion module (Running Suture).

Table I. The five groups of warm-up modules used for curriculum selection.

Simulation Module	Group 1	Group 2	Group 3	Group 4	Group 5
Running Suture	X		X	X	X
Ring and Rail		X	X	X	X
Match Board				X	X
Suture Sponge					X

Video Review

Video was obtained from each criterion (running suture) VR task session, and two blinded expert surgeons reviewed videos of the performances and rated them using the Global Evaluative Assessment of Robotic Skills (GEARS) tool.[30] The 'Autonomy 'domain of the tool was removed as it could not be used with a post-performance video review, leaving a total possible score of 5-25.

Statistical Models

The primary goals of Aim 1 were to evaluate (1) the magnitude of skills decay between the two standardized assessment intervals (monthly and bi-weekly), and (2) the curriculum which optimizes the participants GEARS scores and results in the smallest deviation from performer baseline performance We used a linear mixed effects model which accounts for interval assessment period, timing of assessment, and contrast between the different curricula. This model included a random intercept to account for clustering of study participants, and a fixed effect to adjust for assessment time between different trials. To account for the average percentage of skill decay between groups, the contrast in GEARS scores between groups performing monthly and groups performing bi-weekly was accounted for in the model. The module which yielded the highest average combined GEARS score after accounting for all of these factors was used as the optimal warm-up curriculum for Aim 2.

AIM 2

Subject Recruitment

Aim 1 practicing surgeons were allowed to participate in Aim 2 of the study in addition to newly recruited surgeons.

Randomization & Warm-Up Protocol

Individual scheduled surgeries from each recruited surgeon were randomized to either receive warm-up or not, with each surgeon acting as their own control so that intra-surgeon performance variability did not impact the ability to observe an intervention effect. Surgeons who were randomized to receive the warm-up curriculum did so on the dVSS for either the da Vinci Si or Xi depending on the available simulator. Participating surgeons completed between 1-40 procedures.

OR Data Capture

Endoscopic video and robotic instrument position was captured for each procedure using the Intuitive Surgical dvLogger system which captures data directly from the robots Application Programming Interface (API). Capturing these data synchronized to video allowed for assessing performance with the GEARS and tool motion.

Data Aggregation & Manipulation

All video and tooltip data were captured and performances were parsed into 'SI' or 'XI' labels. Any Personally identifiable information displayed in the videos was censored to create black boxes over these areas.

Video Segmentation

After confirming the timestamps at which the subject surgeon started operating, we obtained the first 15 minutes of surgery performed. The practicing surgeon's complete video was then uploaded as "unlisted" to Youtube for blinded video review

Objective Metric Calculation

The kinematic tooltip data was used to obtain summary metrics based on periods when the surgeon was performing, including rate of orientation change (ROC), spectral arc length (SAL), normalized angular displacement (NAD), mean velocity (MV), laterality (LR), and camera use per minute (CUP); all associated with technical skill.

Video Review

For expert surgeon review, a web-based platform was created to distribute the videos accompanied by the GEARS tool to blinded surgeons and non-expert crowds. Only a subset of videos was used for surgeon review as it became unfeasible to have expert surgeons review over 300 videos.

Each GEARS score was recorded for each performance to analyze the intraclass correlation (ICC) among ratings for the same video. Statistical calculations were computed using R. [32] C-SATS Inc. (Seattle, WA) was used to compile reviews of technical skill for each video, obtained from crowds of non-experts, to offset the limited availability of expert surgeons. The use of crowd-sourced assessment of technical skills (CSATS) has been validated as a rapid means of reliably acquiring expert-like quantitative surgical assessments. [33-37

RESULTS

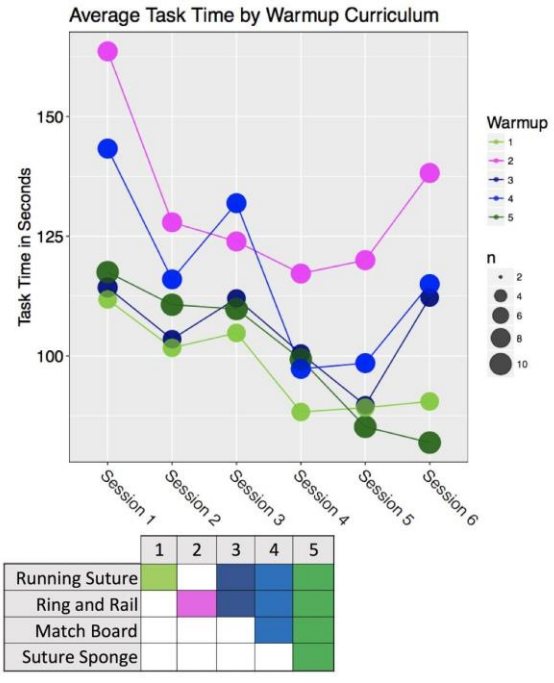
AIM 1

Forty-one participants completed the study in Aim 1, with demographics from these surgeons shown in eTable 2. As can be seen in eFigure 3, the average TT for the different groups had a downward trend in which the surgeons became quicker as they proceeded through sessions. The “Ring and Rail 2” lone module, group 2, was the least effective at lowering TT while also, as seen in eFigure 4, resulted in the lowest GEARS scores. Groups 1 and 5 resulted in the most favorable TT and GEARS scores. Group 1 participants only primed with the “Running Suture 2” module and had a higher GEARS score (2.1 points higher +/- 0.86, $p = 0.02$). Compared to group 2, group 1 was on average 31.3 seconds faster (+/- 14.02 seconds, $p = 0.03$). GEARS scores from groups 1 and 5 were not significantly different from each other ($p = 0.47$ and $p = 0.70$), although group 1 had the least variation in GEARS scores and TT of all groups. “Running Suture 2” which was also the criterion task, thus was chosen as the warm-up curriculum to use before surgical procedures. It had comparable results as Group 5 and practically could be completed in 2-3 minutes

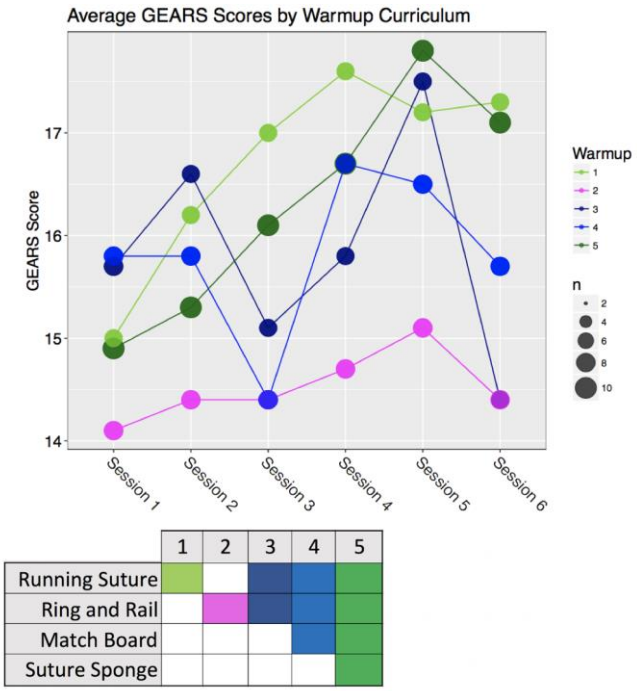
eTable 2 Enrollment and demographics for participants of Aim 1.

	Hospital 1	Hospital 2	Hospital 3/4	Total
Target Enrollment	20	20	30	70
Total Randomized	**5 (25)**	**29 (145)**	**12 (40)**	**46 (66)**
Completed Participation	**3 (15)**	**27 (135)**	**11 (37)**	**41 (59)**
Consented	7 (35)	38 (190)	29 (97)	74 (106)
Gender				
Female	1 (14)	10 (26)	9 (31)	20 (27)
Male	6 (86)	28 (74)	17 (59)	51 (69)
Other	0 (0)	0 (0)	0 (0)	0 (0)
Seniority				
Attending	7 (100)	12 (32)	3 (10)	22 (30)
Senior	0 (0)	9 (24)	9 (31)	18 (24)
Junior	0 (0)	17 (45)	14 (48)	31 (42)
Specialty				
Urology	3 (43)	12 (32)	12 (41)	27 (36)
Gynecology	1 (14)	9 (24)	2 (7)	12 (16)
General	2 (29)	8 (21)	10 (34)	20 (27)
Thoracic	0 (0)	0 (0)	0 (0)	0 (0)
Cardiothoracic	0 (0)	0 (0)	0 (0)	0 (0)
ENT	1 (14)	9 (24)	2 (7)	12 (16)
Handedness				
Left	1 (14)	1 (3)	1 (3)	3 (4)
Right	5 (71)	35 (92)	23 (79)	63 (85)
Ambidextrous	1 (14)	2 (5)	2 (7)	5 (7)
Musical Experience	2 (29)	22 (58)	19 (66)	43 (58)
Deployed	0 (0)	7 (18)	0 (0)	7 (9)

eFigure 4 Average task time for each of the 5 warm-up modules in aim 1. Running Suture, group 1, had the largest decrease in average task time.



eFigure 5 Average GEARS score for each of the 5 warm-up modules in aim 1. Running Suture, group 1, had the largest increase in average GEARS score.



	1	2	3	4	5
Running Suture					
Ring and Rail					
Match Board					
Suture Sponge					

AIM 2

After identifying the optimal VR warm-up curriculum, forty practicing surgeons were enrolled in Aim 2, with 34 surgeons having completed at least one surgery during the study. Demographics questionnaire results are shown in the eTable 3. Video of 432 robot-assisted surgeries were performed and recorded. Kinematic data failed to capture in 89 cases leaving 343 surgeries with both videos and corresponding tooltip position kinematic data. The amount of each procedure as a percentage of total sessions is shown in eTable 4.

eTable 3 Enrollment and demographics information for Aim 2 of the study, including handedness, surgical specialty, as well as musical experience and deployment history.

	Hospital 2	Hospital 3	Total
Sessions with Video/Kinematic	17	326	343
Completed Sessions	41	391	432
Consented	12 (200)	22 (110)	34 (131)
Deployed Currently	0 (0)	0 (0)	0 (0)
Lost to Follow-Up	0 (0)	0 (0)	0 (0)
Gender			
Female	1 (8)	8 (36)	9 (26)
Male	10 (83)	11 (50)	21 (62)
Other	0 (0)	0 (0)	0 (0)
Specialty			
Urology	2 (17)	8 (36)	10 (29)
Gynecology	0 (0)	9 (41)	9 (26)
General	6 (50)	1 (5)	7 (21)
Thoracic	0 (0)	0 (0)	0 (0)
Cardiothoracic	1 (8)	0 (0)	1 (3)
Handedness			
Left	0 (0)	0 (0)	0 (0)
Right	11 (92)	19 (86)	30 (88)
Ambidextrous	0 (0)	0 (0)	0 (0)
Musical Experience	3 (25)	6 (27)	9 (26)
Deployed Previously	9 (75)	0 (0)	9 (26)

eTable 4 Percentage of total cases for each of the more popular robot-assisted procedures completed during the study.

	Percentage of Total Sessions
Procedure Type	
Prostatectomy	25.7%
Hysterectomy	23.8%
Partial Nephrectomy	10.0%
Liver Resection	2.8%
Nissen Fundoplication	1.2%
Partial Colectomy	0.2%
Other	36.3%

GEARS Review Outcome

Ratings by seven expert surgeons were obtained for a subset of 45 videos, with two raters rating all videos and the other five rating from 7 to 44 videos each. ICC for expert ratings was 0.38(95% CI 0.25 to 0.54), indicating poor agreement. When the two raters with the greatest numbers of missing observations were removed, ICC was 0.45(95% CI 0.30 to 0.60), still indicating poor agreement. ICC estimates and their confidence intervals were based on an individual rating, consistency, 2-way random-effects model for incomplete datasets.[38]

We assessed the effect of the warm-up on GEARS score, our primary outcome, using both a linear mixed model with a random intercept for each surgeon ($p = 0.98$) and a nonparametric modified Friedman test ($p = 0.34$), with both finding no significant difference between cases receiving or not receiving the warm-up (Table 2). Robot type (XI vs SI) was also not significantly related to GEARS score ($p = 0.24$), nor was any robot by treatment interaction detected ($p = 0.67$, Figure 1).

Figure 1 Effect of warm-up for each type of robot used in the study, Xi and Si DaVinci robots.

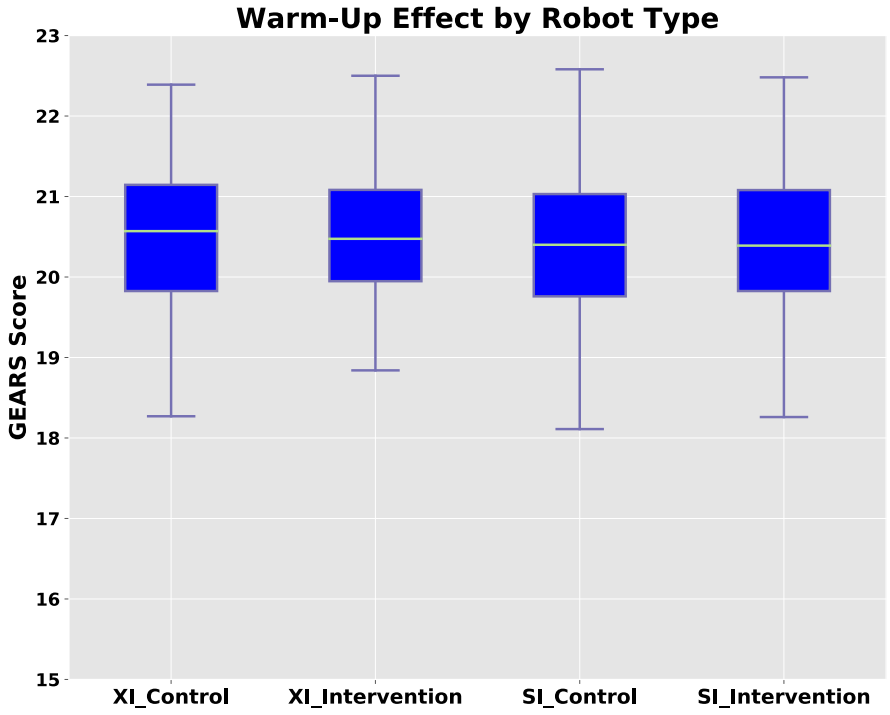


Table 2 Results from statistical tests on the kinematic and event measures recorded, for MV (mean velocity), NAD (normalized angular displacement), SAL (spectral arc length), CUP (camera use per minute), LR (laterality), and ROC (rate of orientation change). Both tests show poor results.

(a) Mixed model results for warm-up effect on kinematic and event outcomes.

Measure	Est. β (p-value)
MV	0.06 (0.40)
NAD	-0.06 (0.43)
SAL	0.06 (0.66)
CUP	0.03 (0.77)
LR	-0.01 (0.79)
ROC	0.02 (0.89)

(b) Modified Friedman test results for warm-up effect on kinematic and event outcomes.

Measure	p-value
MV	0.35
NAD	0.29
SAL	0.31
CUP	0.68
LR	0.49
ROC	0.16

OBJECTIVE METRICS OUTCOME

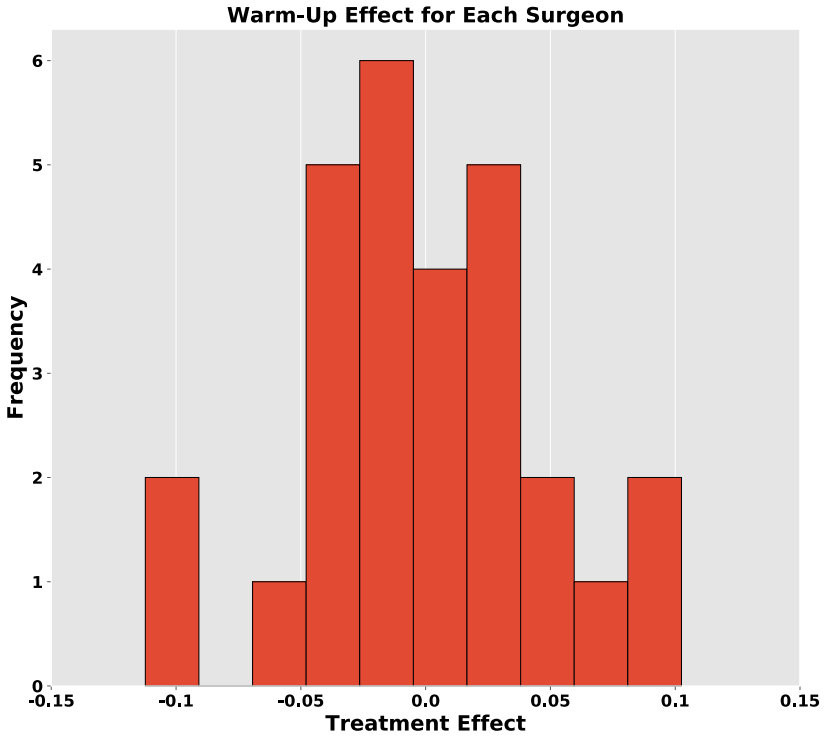
We also examined the effect of the warm-up on six kinematic metrics MV, ROC, NAD, SAL, CUP, LR using both linear mixed models and a nonparametric modified Friedman test, finding no significant differences between treatment arms. We used the Holm-Sidak procedure to maintain familywise alpha at .05. Kinematic and event outcomes were not significantly associated with GEARS score as quantified by repeated measures correlation (using R package rmcorr [39], results shown in eTable5).

eTable 5 Correlation of the kinematic and event measures to GEARS scores by repeated meausred correlation. These measures were not significantly associated with GEARS scores.

	Correlation (+/- 95% CI)	p-value
MV	0.08 (0.11)	0.16
ROC	-0.01 (0.11)	0.81
NAD	-0.06 (0.11)	0.32
SAL	-0.02 (0.11)	0.69
LR	-0.01 (0.11)	0.86
CUP	0.08 (0.11)	0.16

A linear mixed model including both a random intercept and a random slope for the effect of the treatment on each surgeon was used to quantify the range of effects the warm-up had on different surgeons. Treatment effects for individual surgeons on GEARS score ranged from -.10 to .09 points, which we judged not to be clinically meaningful, thus precluding the possibility of identifying subgroups of surgeons who benefitted from the warm-up to any practically significant degree (Figure 2).

Figure 2 A clear normal curve appears from the analysis of warm-up effect for each individual surgeon. This shows strong evidence that the VR warm-up module did not affect technical skills proficiency.



DISCUSSION

No statistically significant difference in technical skill was observed when surgeons performed a virtual reality warm-up and when they did not. Tests controlling for individual surgeons, Si vs. Xi, and time since last surgical performance found no evidence that performing a warm-up curriculum prior to surgery elevated surgical readiness. The null findings were unexpected since we had seen a warm-up benefit is our previous dry lab study [19] and the volume of data and trial sessions suggest that these findings are valid. Sports medicine research concludes that performing a warm-up through the use of proprioceptive neuromuscular facilitation techniques provides improvements to athletic ability, and can enhance skill in athletics. [40, 41] A study measuring brain function of young athletes performing a warm-up prior to cycling training found that warm-ups resulted in a significantly increased amount of oxyhemoglobin levels in the prefrontal cortex and motor areas of the brain [42], which is also the case for general motor activity and has been shown to result in shorter reaction times. [43,44] Other papers on motor learning and the brain suggest the prefrontal cortex and anterior cingulate cortex are in control of the initial phases of motor skills learning [45,46], as well as a study finding behavioral modification to lead to activation in these regions.[47]

These same areas of the brain, related to visual imagery and motion processing, have been found to be activated during surgical performance training tasks [44,48] suggesting that some form of training or warm-up could potentially aid surgeons in preparing for surgery. A study which included surgeons of mixed experience levels concluded that for less experienced surgeons, a larger modification in prefrontal cortex activation appeared to confer technical performance adaptation, suggesting this warm-up effect may only be present in novice surgeons.[49]

Past studies have shown differing results from the use of a warm-up prior to surgical procedures. A study performed with residents using laparoscopic procedures found that brief warm-ups performed prior to surgery significantly improved intraoperative performance, regardless of case difficulty [50] Alternatively, some studies have shown no warm-up effect. A study of nine laparoscopic surgeons performing either no warm-up, laparoscopic trainer box warm-up, or PlayStation 2 warm-up, showed that out of 75 recorded cases, warm-up prior to surgery was not essential to performance, as measured by an objective assessment tool.[51] A randomized, controlled study in which surgeons performed VR warm-up training before laparoscopic cases found that overall performing the warm-up did not increase psychomotor skills during surgery as measured by global assessment scores, and also did not influence operating time or complication rates.[52] We previously have shown that in dry lab settings of robot-assisted procedures in which surgeons performed VR warm-up prior to criterion tasks such as suturing, warm-up yielded a significant reduction in errors. Moreover, the study found that expert surgeons experienced a significant improvement from warm-up in their task time and economy of motion, while these findings did not extend to less-experienced surgeons.[19]

This research study had limitations. Surgeons performed different types of procedures which consist of several different types of movements and skills needed to complete. Despite all surgeons enrolled being faculty surgeons, varying degrees of skill existed. Although it is most likely the first 15 minutes of surgery were the most likely to derive a warm-up benefit, what happens during this time may be relatively simple steps in a case, which would not tax the performer as much as the steps to follow. The VR module used may have not primed surgeons enough to boost arousal or performance to a significant degree. Additionally, these modules may stimulate important centers of the brain that confer good surgical performance but more so in the domain of surgical decision-making than of technical skill, and this was not studied. This study looked at intra-operative performance only and was not designed to assess patient outcomes, thus understanding whether warm-up had an impact on success of surgery could not be ascertained. Other aspects of the beginning of a surgical case including port access, conventional laparoscopic manipulation, and patient set-up may have already primed surgeons who did not warm-up. Assessing these factors was beyond the scope of this study. Warm-up has been shown to have a benefit in conventional laparoscopy intra-operatively [18,53], and it's possible that RAS may not be as technically challenging, thus not requiring a technical skills warm-up.

Pre-operative warm-up may take the form of mental visualization, reviewing patient images beforehand, describing steps of a case to a trainee, patient positioning, placement of robotic ports,

and initial conventional laparoscopic tissue manipulation. These are all important and may have a significant positive impact on a surgeon's priming such that assessment of technical skills alone may not have detected the added value of VR warm-up.

CONCLUSIONS

Counter to the majority of existing literature in conventional laparoscopic surgery, and observations

made by this author group in a RAS dry lab setting, performing a VR warm-up curriculum prior to robot-assisted human surgery does not significantly improve the technical skill of a practicing surgeon. Because surgery is a dynamic process requiring thoughtful choreography before, during and after a case, isolating a specific pre-operative surgeon behavior or effort that may impact the surgeon's readiness may be difficult.

ACKNOWLEDGEMENTS

Thank you to each of the faculty surgeons for giving their time to provide us with video review for several videos of robotic surgery procedures: Ardavan Akhavan, Walid Farhat, Christopher Weight, Jonathan Ellison, Tim Brand, Thomas Lendvay, and Matthew Sorensen. Anthony Jarc from Intuitive Surgical provided appreciated technical support during the data processing phase of the project. Additionally, we'd like to thank all the surgeons and residents who graciously participated in this study and committed so much of their time.

REFERENCES

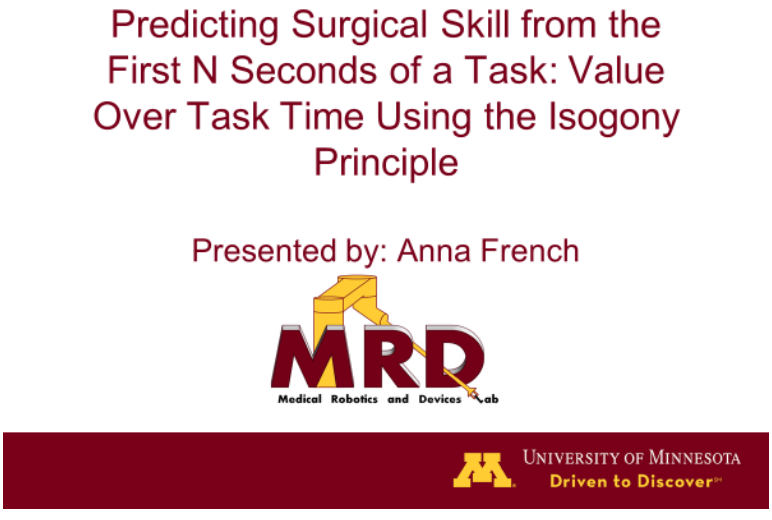
1. Kohn L, JM C, Donaldson M, eds. To Err is Human: Building a Safer Health Care System. Washington, DC: National Academy Press, 2000.
2. Zhan C, Miller M. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. JAMA 290:1868, 2003.
3. Murphy S, Xu J, Kochanek K. Deaths: Preliminary data for 2010. National Vital Statistics Reports 60(4), 2012, http://www.cdc.gov/nchs/data/nvsr60/nvsr60_04.pdf.
4. Shreve J, of Actuaries Health Section S, (Firm) M. The economic measurement of medical errors. Society of Actuaries; 2010, http://soa.org/Files/Research/Projects/research-econ-measurement.pdf.
5. Sroka G, Feldman L, Vassiliou M, et al. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room - a randomized controlled trial. Am J Surg. 199:115, 2010.
6. Stefanidis D, Korndorffer JR Jr, Sierra R, Touchard C, Dunne JB, Scott DJ. Skill retention following proficiency-based laparoscopic simulator training. Surgery 138: 165-170, 2005.
7. Perez RS, Skinner A, Weyhrauch P, Nehaus J, Lathan C, Schwaitzberg SD, Cao CG. Prevention of surgical skill decay. Military Medicine 178:588-92, 2013.
8. Sheetz KH, Claflin J, Dimick JB. Trends in the Adoption of Robotic Surgery for Common Surgical Procedures. JAMA Network Open 3(1): 2020.
9. Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat A-HS, Dellinger EP, Herbosa T, Joseph S, Kibatala PL, Lapitan MCM, Merry AF, Moorthy K, Reznick RK, Taylor B, Gawande AA. A Surgical safety checklist to reduce morbidity and mortality in a global population. NEJM 360: 491-499, 2009.
10. Nascon, J, Schmidt, RA. The activity-set hypothesis for warm-up decrement. Journal of Motor Behavior, 3:1-16, 1971.
11. Wrisberg, CA, Salmoni, AW, Schmidt, RA, Warm-up effects in the learning of discrete motor skills. Acta Psychologica 29:311-320, 1975.
12. Anshel, MA. the effect of arousal on warm-up decrement. Research Quarterly for Exercise and Sport 56:1-9, 1985.
13. Anshel, MH, Wrisberg, CA. Reducing warm-up decrement in the performance of the tennis serve. J Sport & Exercise Psychology 15:290-303, 1993.
14. Wetter, PA. Resubmission of gap analysis workshop for training for reintegration of surgical skills. Proceedings of the 2011 Society of Laparoendoscopic Surgeons Conference. Miami, FL, 2011.

15. Arthur W Jr, Winston B Jr, Stanush PL, McNelly TL. Factors that influence skill decay and retention: A quantitative review and analysis. Human Performance 11:57-101, 1998.
16. Seymour NE, Gallagher AG, Roman SA, O' Brien MK, Bansal VK, Andersen DK, Satava RM. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 236:458-63, 2002.
17. Kahol K, Satava R, Ferrara J, Smith ML. Effect of short-term pretrial practice on surgical proficiency in simulated environments: a randomized trial of the "preoperative warm-up" effect. JACS 208:255, 2009.
18. Calatayud D, Arora S, Aggarwal R, Kruglikova I, Schulze S, Funch-Jensen P, Grantcharov T. Warm-up in a virtual reality environment improves performance in the operating room. Ann Surg. 251: 1181, 2010.
19. Lendvay TS, Brand TC, White L, Kowalewski T, Jonnadula S, Mercer L, Khorsand D, Andros J, Hannaford B, Satava RM. Virtual reality robotics surgery warm-up improves surgical performance: A prospective randomized controlled study. JACS 216:1181-1192,2013.
20. Stefanidis D, Korndorffer JR Jr, Sierra R, Touchard C, Dunne JB, Scott DJ. Skill retention following proficiency-based laparoscopic simulator training. Surgery 138:165-170, 2005.
21. Mimic Technologies Inc. Da Vinci Skills Simulator (dvss). 2018.
22. Holst D, Kowalewski T, White L, Brand T, Harper J, Sorensen M, Kirsch S, and Lendvay TS. Crowd-Sourced Assessment of Technical Skills: An Adjunct to Urology Resident Surgical Simulation Training. Journal of Endourology, 29(5): 604-609, 2015.
23. Kowalewski T, Comstock B, Sweet R, Schaffhausen C, Menhadji A, Averch T, Box G, Brand T, Ferrandino M, Kaouk J, Knudsen B, Landman J, Lee B, Schwartz BF, McDougall E, and Lendvay TS. Crowd-sources assessment of technical skills for validation of basic laparoscopic urologic skills tasks. The Journal of Urology, 195(6): 1859-1865, 2016.
24. Chen J, Oh PJ, Chang N, Shah A, Montez J, Jarc A, Guo L, Gill IS, and Hung AJ. Use of Automated Performance Metrics to Measure Surgeon Performance during Robotic Vesicourethral Anastomosis and Methodical Development of a Training Tutorial. Journal of Urology 200(4): 895-902, 2018.
25. Nguyen JH, Chen J, Mashall SP, Ghodoussipour S, Chen A, Gill IS, and Hung AJ. Using objective robotic automated performance metrics and task-evoked pupillary response to distinguish surgeon expertise. World Journal of Urology, 2019.
26. Balasbramanian S, Melendez-Calderon A, Burdet E. A Robust and Sensitive Metric for Quantifying Movement Smoothness. IEEE Transactions of Biomedical Engineering 59(8), Aug. 2012.
27. Kowalewski TM, White LW, Lendvay TS, Jiang IS, Sweet RS, Wright A, Hannaford B, and Sinanan MN. Beyond task time: automated measurements augments fundamentals of laparoscopic skills methodology. Journal of Surgical Research 192(2): 329-338, 2014.
28. Sharon Y, Lendvay TS, and Nisky I. Instrument Orientation Metrics for Robot-Assisted and Open Surgical Skill Evaluation. IEEE Transactions on Human-Machine Systems, 2017.
29. Hung AJ, Zehnder P, Patil MB, Cai J, Ng CK, Aron M, Gill IS, Desai MM. Face, content and construct validity of a novel robotic surgery simulator. J Urol 186: 1019-1025, 2011.
30. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. Journal of Urology 187:247-252, 2012.
31. Python Software Foundation. Python Language Reference, version 3.6. Available at http://www.python.org.
32. R Core Team (2018) R: A Language and Environment for Statistical Computing, Vienna, Austria. http://www.R-project.org.
33. Chen SP, Kirsch S, Zlatev DV, Chang T, Comstock B, Lendvay TS, Liao JC. Optical Biopsy of Bladder Cancer Using Crowd-Sourced Assessment. JAMA surgery 151(1):90-3, 2016.
34. Lendvay TS, White L, Kowalewski T. Crowdsourcing to assess surgical skill. JAMA Surgery 150(11): 1086-7, 2015.
35. Chen SP, Kirsch S, Zlatev DV, Chang T, Comstock B, Lendvay TS, Liao JC. Optical Biopsy of Bladder Cancer Using Crowd-Sourced Assessment. JAMA Surgery 151(1):90-93, 2016.
36. Holst D, White L, Brand T, Harper J, Lendvay T. Crowd-Sourced Assessment of Technical Skills (C-SATS): An Adjunct to Urology Resident Simulation Training. American Urological Association Annual Meeting, Orlando, FL, May 16-21, 2014.
37. White L, Kowalewski T, Brand T, Harper J, Comstock B, Hannaford B, Lendvay T. Crowd-Sourced Assessment of Technical Skills (C-SATS): Fast, Economical, and Accurate assessment of Robotic Surgery. American Urological Association Annual Meeting, Orlando, FL, May 16-21, 2014.
38. Coefficients of Interrater Reliability - Generalized for Randomly Incomplete Datasets. Brueckl M and Heuer F, 2018. https://cran.r-project.org/web/packages/irrNA/index.html
39. Repeated Measures Correlation. Jonathan Z. Bakdash and Laura R. Marusich, 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5383908/
40. Shellock FG, Prentice WE. Warming-Up and Stretching for Improved Physical Performance and Prevention of Sports-Related Injuries. Sports Medicine 2: 267-278, 1985.
41. Daneshjoo A, Mokhtar AH, Rahnama N, and Yusof A. Effects of the 11+ and Harmoknee Warm-Up Programs on Physical Performance Measures in Professional Soccer Players. J Sports Sci Med 12(3): 489-496, 2013.
42. Techayusukcharoen R, Iida S, Aoki C. Observing brain function via functional near-infrared spectroscopy during cognitive program training (dual task) in young people. Journal of Physical Therapy Science 31(7), 2019.

43. Maki A, Yamashita Y, Ito Y, Watanabe E, Mayanagi Y, and Koizumi H. Spatial and temporal analysis of human motor activity using noninvasive NIR topography. Med Phys 22(12): 1997-2005, 1995.
44. Leff DR, Orihuela-Espina F, Atallah L, Athanasiou T, Leon JJH, Darzi AW, and Yang GZ. Functional prefrontal reorganization accompanies learning-associated refinements in surgery: A manifold embedding approach. Computer Aided Surgery 13(6): 325-339, 2008.
45. Kelly AM, and Garavan H. Human functional neuroimaging of brain changes associated with practice. Cereb Cortex 15(8): 1089-1102, 2005.
46. Halsband U, and Lange RK. Motor learning in man: a review of functional and clinical studies. J Physiol Paris 99(4-6): 414-424, 2006.
47. Walton MR, Croxson PL, Behrens TEJ, Kennerley SW, and Rushworth MFS. Adaptive decision making and value in the anterior cingulate cortex. NeuroImage 36(2): 2007.
48. Wanzel KR, Anastakis DJ, McAndrews MP, Grober ED, Sidhu RS, Taylor K, Mikulis DJ, and Hamstra SJ. Visual-spatial ability and fMRI cortical activation in surgery residents. Am. J. Surg. 193: 507-510, 2007.
49. Leff DR, Elwell CE, Orihuela-Espina F, Atallah L, Delpy DT, Darzi AW, and Yang GZ. Changes in prefrontal cortical behaviour depend upon familiarity on a bimanual coordination task: An fNIRS study. NeuroImage 39(2): 805-813, 2008.
50. Chen CCG, Green IC, Colbert-Getz Jm, Steele K, Chou B, Lawson SM, Andersen DK, and Satin AJ. Warm-up on a simulator improves residents' performance in laparoscopic surgery: a randomized trial. International Urogynecology Journal 24: 1615-1622, 2013.
51. Weston MK, Stephens JH, Schafer A, and Hewett PJ. Warm-up before laparoscopic surgery is not essential. Anz Journal of Surgery 84(3): 143-147, 2014.
52. Polterauer S, Husslein H, Kranawetter m, Schwamels R, Reinthaller A, Heinze G, and Grimm C. Effect of Preoperative Warm-up Exercise Before Laparoscopic Gynecological Surgery: A Randomized Trial. Journal of Surgical Education 73(3): 429-432, 2016.
53. Lee JY, Mucksavage P, Kerbl DC, Osann KE, Winfield HN, Kahol K, McDougall EM. Laparoscopic Warm-up Exercises Improve Performance of Senior-level Trainees During Laparoscopic Renal Surgery. JEndourol26(5):545-550,2012.
54.

Appendix I

"Predicting Surgical Skill from the First N Seconds" at the IPCAI conference, Barcelona, Spain June 2017



Objectives

- Investigate task-blind skill classification
- Investigate skill classification prior to task completion



Dataset

Electronic Data Generation and Evaluation (EDGE)

- 67 novices, 157 experts
- Each trial accompanied by video and tool motion data recorded at 30Hz along with FLS score for trial



Feature Design



$$v = \gamma \kappa^{-\frac{1}{3}}$$

Lacquaniti et al. 1983

Feature Design



$$v = \gamma\kappa^{-\frac{1}{3}}$$



$$\gamma = v\kappa^{\frac{1}{3}}$$

Feature Design



$$v = \gamma\kappa^{-\frac{1}{3}}$$



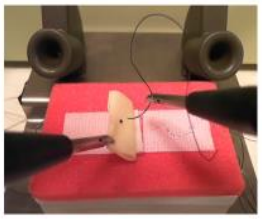
$$\gamma = v\kappa^{-\frac{1}{3}}$$

γ [illegible] $\kappa^{\frac{1}{3}}$

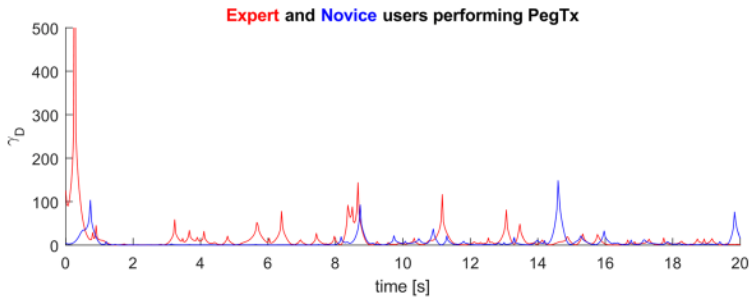
Feature Design

$$\gamma_d = v_d \circ \kappa_d^{-\frac{1}{3}}$$

$$\gamma_{nd} = v_{nd} \circ \kappa_{nd}^{-\frac{1}{3}}$$



Feature Design



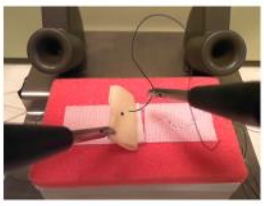
Feature Design

$$\gamma_d = v_d \circ \kappa_d^{-\frac{1}{3}}$$

$$\gamma_{nd} = v_{nd} \circ \kappa_{nd}^{-\frac{1}{3}}$$

$\sigma(\gamma_d)$

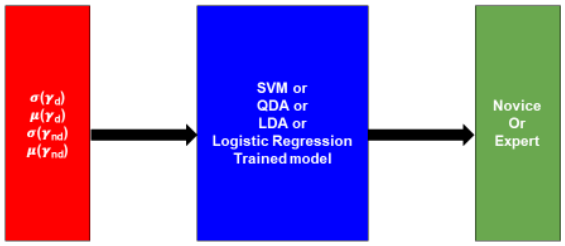
$\mu(\gamma_d)$

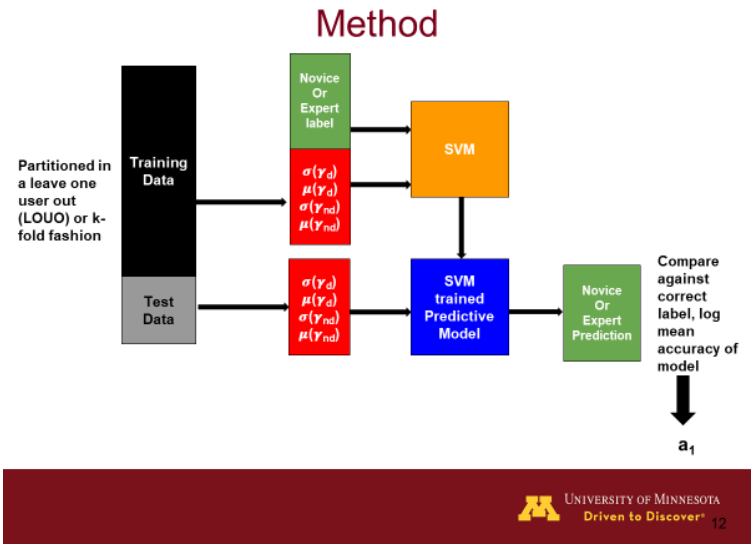


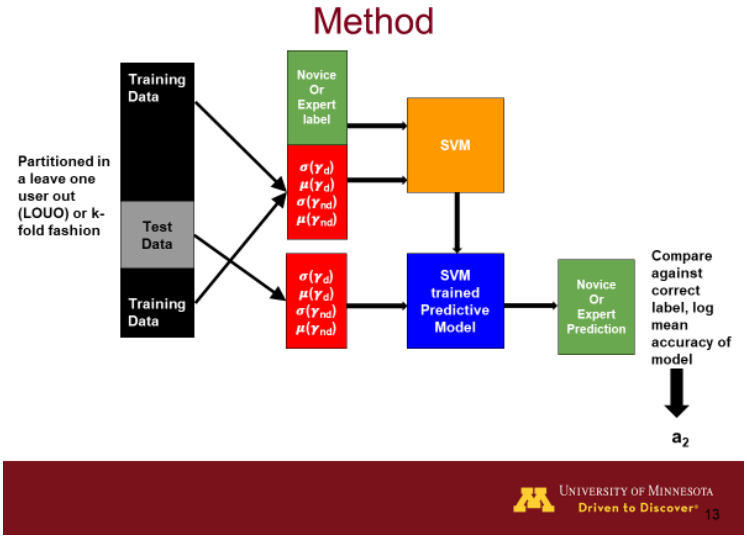
$\sigma(\gamma_{nd})$

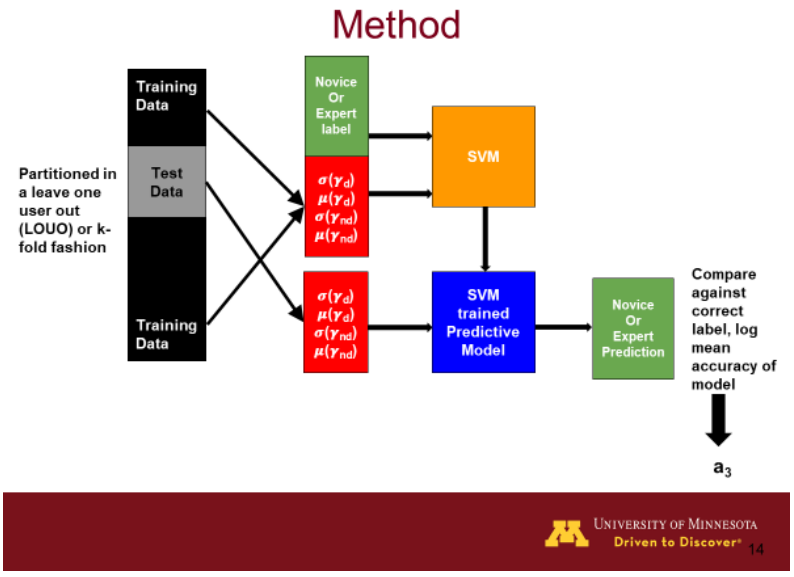
$\mu(\gamma_{nd})$

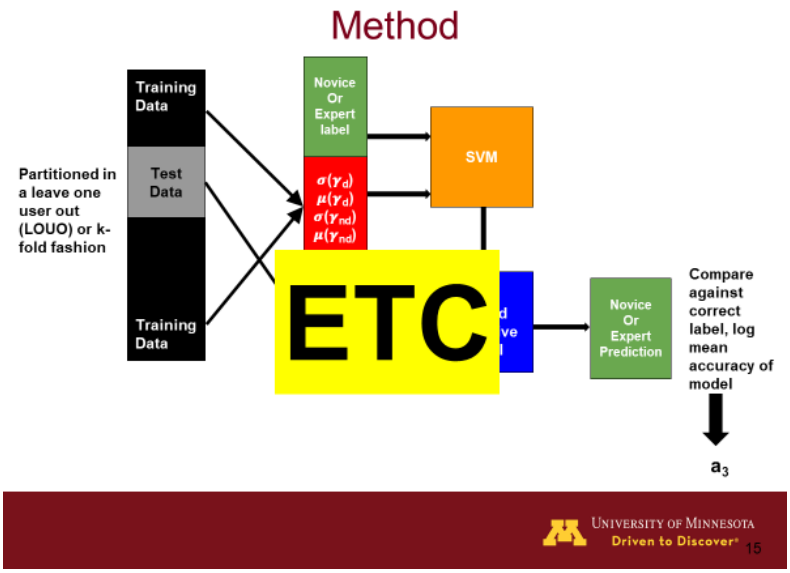
Hypothesis: $\sigma(\gamma_d)$, $\mu(\gamma_d)$,$\sigma(\gamma_{nd})$, $\mu(\gamma_{nd})$ features can predict FLS skill level

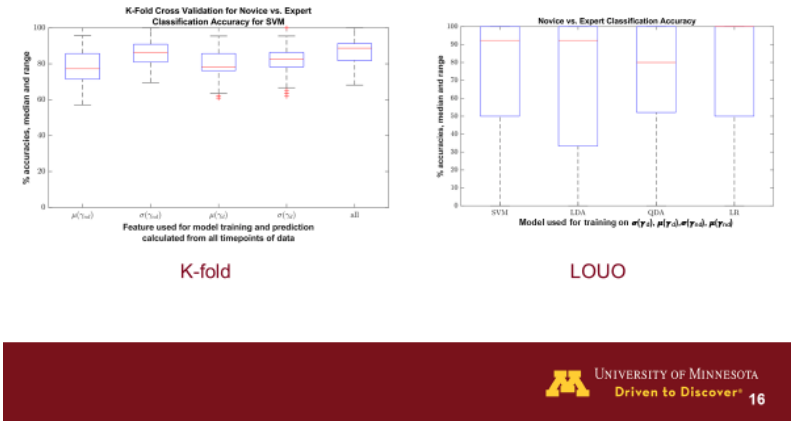



Method
Partitioned in a leave one user out (LOUO) or k-fold fashion
Training Data
Test Data
Novice Or Expert label
$\sigma(\gamma_d)$ $\mu(\gamma_d)$ $\sigma(\gamma_{nd})$ $\mu(\gamma_{nd})$
SVM
SVM trained Predictive Model
Novice Or Expert Prediction
Compare against correct label, log mean accuracy of model
a_1
University of Minnesota Driven to Discover 12


Method
Partitioned in a leave one user out (LOUO) or k-fold fashion
Training Data
Test Data
Training Data
Novice Or Expert label
$\sigma(\gamma_d)$ $\mu(\gamma_d)$ $\sigma(\gamma_{nd})$ $\mu(\gamma_{nd})$
SVM
SVM trained Predictive Model
Novice Or Expert Prediction
Compare against correct label, log mean accuracy of model
a_2
University of Minnesota Driven to Discover 13

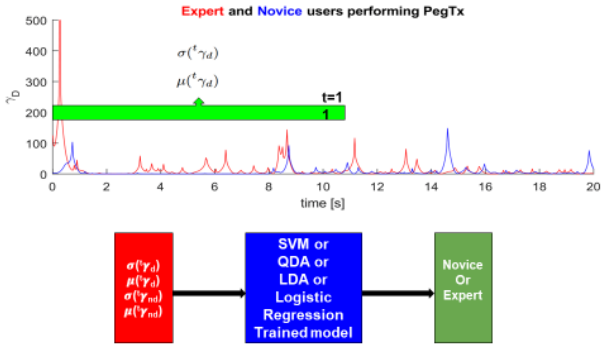

Method
Partitioned in a leave one user out (LOUO) or k-fold fashion
Training Data
Test Data
Training Data
Novice Or Expert label
$\sigma(\gamma_d)$ $\mu(\gamma_d)$ $\sigma(\gamma_{nd})$ $\mu(\gamma_{nd})$
SVM
SVM trained Predictive Model
Novice Or Expert Prediction
Compare against correct label, log mean accuracy of model
a_3
University of Minnesota Driven to Discover 14

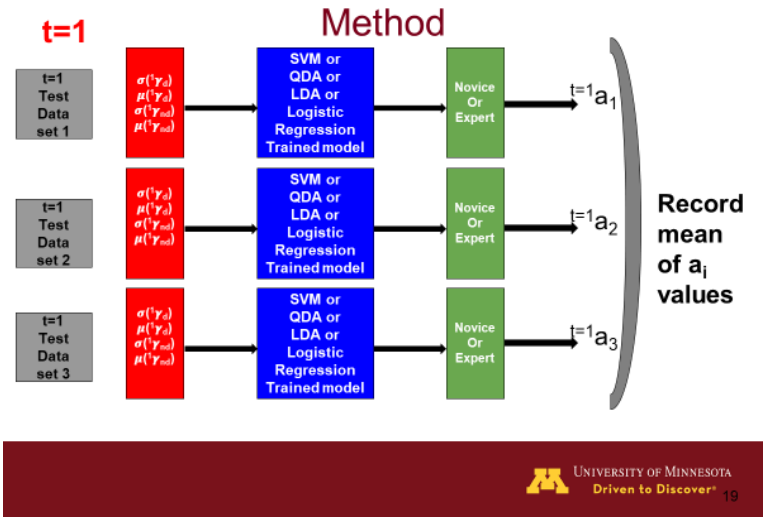

Method
Training Data
Test Data
Training Data
Partitioned in a leave one user out (LOUO) or k-fold fashion
Novice Or Expert label
$\sigma(\gamma_d)$
$\mu(\gamma_d)$
$\sigma(\gamma_{nd})$
$\mu(\gamma_{nd})$
SVM
ETC
Novice Or Expert Prediction
Compare against correct label, log mean accuracy of model
a_3
UNIVERSITY OF MINNESOTA
Driven to Discover℠
15

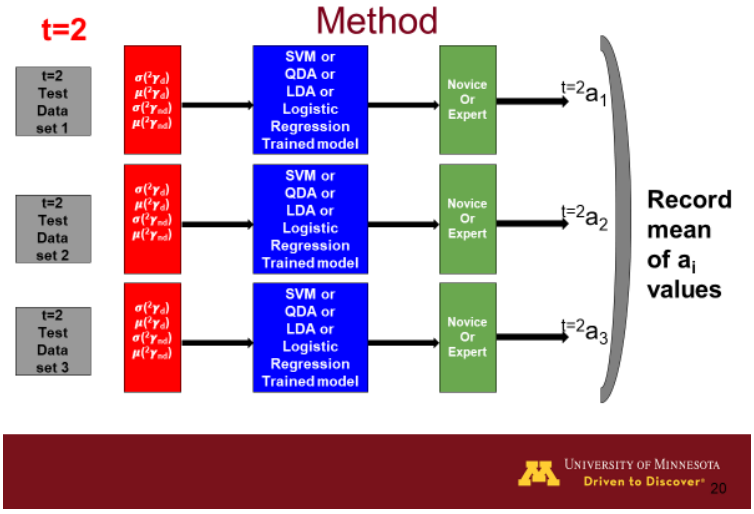

Results
K-Fold Cross Validation for Novice vs. Expert Classification Accuracy for SVM
% accuracies, median and range
Feature used for model training and prediction calculated from all timepoints of data
K-fold
Novice vs. Expert Classification Accuracy
% accuracies, median and range
SVM
LDA
QDA
LR
Model used for training on
LOUO
UNIVERSITY OF MINNESOTA
Driven to Discover℠
16

Hypothesis: Classification can be made Prior to Task Completion

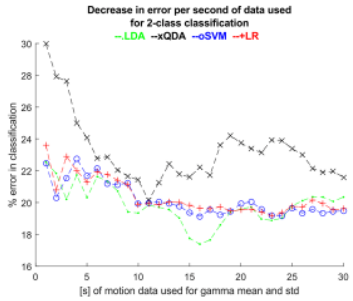
Method





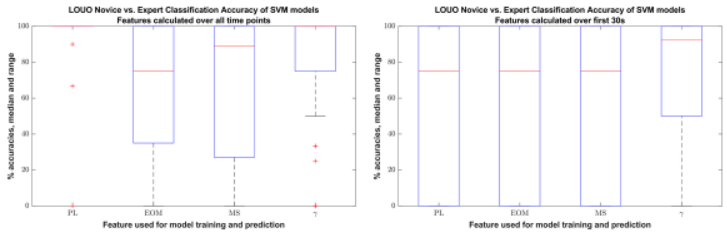


Results: features calculated prior to task end time



Accuracies calculated from K-fold

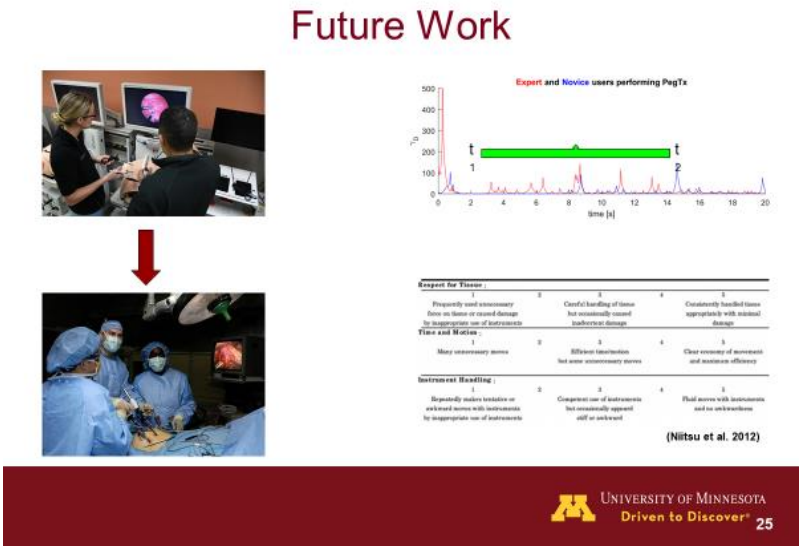
Results: Comparison to baseline features



Full Time Duration Features **First 30s Features**

Conclusion

- Novices and experts can be distinguished with 85% median accuracy using SVM trained models from features $\sigma(\gamma_d)$, $\mu(\gamma_d)$,$\sigma(\gamma_{nd})$, $\mu(\gamma_{nd})$
- Classification between novice and experts can be made within the first 10 seconds of task time.



Appendix J

Abstract for podium at the Society of Pediatric Urology meeting Sept 14-16, 2018, Atlanta

Robotic Surgery Readiness: What is the Optimal Warm-up for Robotic Skills?

Author Block Thomas S. Lendvay, MD, FACS[1], Timothy Brand, MD[2], Anna French, PhD[3], Lois Meryman, BS[4], Nancy Organ, MS[4], Mathew Sorensen, MD, MS[5], Jason Kelly, BS[3], Evalyn George, BS[6], Timothy Kowalewski, PhD[3].
[1]Seattle Children's Hospital, Seattle, WA, USA, [2]Madigan Army Medical Center, Tacoma, WA, USA, [3]University of Minnesota, Minneapolis, MN, USA, [4]University of Washington, Seattle, WA, USA, [5]Puget Sound Veteran Affairs Medical Center, Seattle, WA, USA, [6]Madigan Army Medical Center, Henry M. Jackson Foundation, Tacoma, WA, USA.

Abstract:
BACKGROUND: A surgeon's technical performance directly impacts patient outcomes. Virtual Reality(VR) warm-up, mental rehearsal, and patient/procedure-specific practice may improve surgeon performance. Yet, there is minimal literature on the optimal warm-up for robotic surgery practice. We sought to identify the optimal virtual reality curriculum that would prime (warmed-up) practicing surgeons prior to doing a VR criterion suturing task on the da Vinci Skills Simulator(dVSS). This curriculum will then be used to prime surgeons prior to performing clinical robotic surgery.
METHODS: After IRB approval, surgical trainees and faculty from urology, general surgery, gynecology, and otolaryngology at three medical centers completed the study. Upon recruitment, each subject went through proficiency training on VR skills modules exercising different techniques (suturing, use of the fourth arm, object transfer, retraction). Proficiency benchmarks were derived from two expert robotic surgeons performing these modules. Once proficient, each subject was randomized to one of 5 groups to test which series of warm-up modules optimally prime the surgeon[Figure1]. Each subject performed six trial sessions. Criterion performance metrics included Task Time(TT) and expert video review using Global Evaluative Assessment of Robotic Skills(GEARS), a validated technical skills assessment tool. The randomization group yielding the highest GEARS scoring and best TT dictated the optimalwarm-up curriculum.
RESULTS: Forty-four subjects completed the study. Based on TT and GEARS scores, there were two groups(1 and 5) who demonstrated optimal performance priming - one that included Continuous Suturing, Ring and Rail, Match Board, and Suture

Sponge modules and one that only contained the Continuous Suturing module[Figure2]. Compared to Group 2 which had the worst performance, Group 1 subjects had a higher GEARS score(2.1 points higher +/- 0.86, p = 0.02)[Figure3]. Compared to Group 2 which had the slowest TT, Group 1 was on average 31.3 seconds faster (+/- 14.02 seconds, p = 0.03)[Figure4]. GEARS scores (p = 0.47) and TT(p = 0.70) from Groups 1 and 5 were not significantly different from each other. Group 1 had the least variation(SD) in GEARS scores and TT of all groups, thus was chosen as the optimal VR warm-up.

CONCLUSIONS: This study identified an optimal warm-up curriculum to prime surgeons for VR suturing using the dVSS. We demonstrated that GEARS scores cross-validated with TT. This optimal VR warm-up curriculum can now be deployed into the operating room to test the hypothesis that VR robotic warm-up can prime surgeons for performing clinical robotic surgery.

Funded by DoD TATRC Award#W81XWH-15-2-0300(PI:Lendvay)

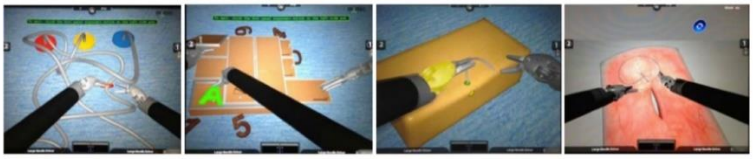


Figure 1. Virtual reality simulator modules. [Ring and Rail, Match Board, Suture Sponge, Running Suture]

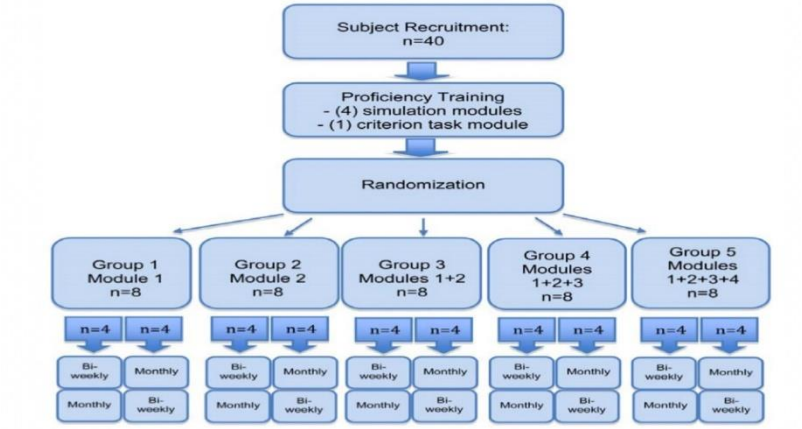


Figure 2. Subject Flow Diagram.

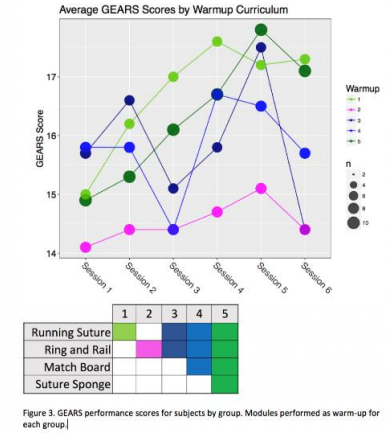


Figure 3. GEARS performance scores for subjects by group. Modules performed as warm-up for each group.

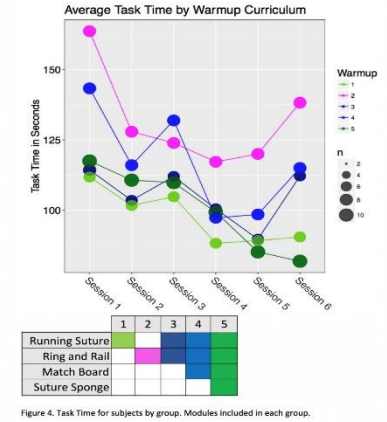
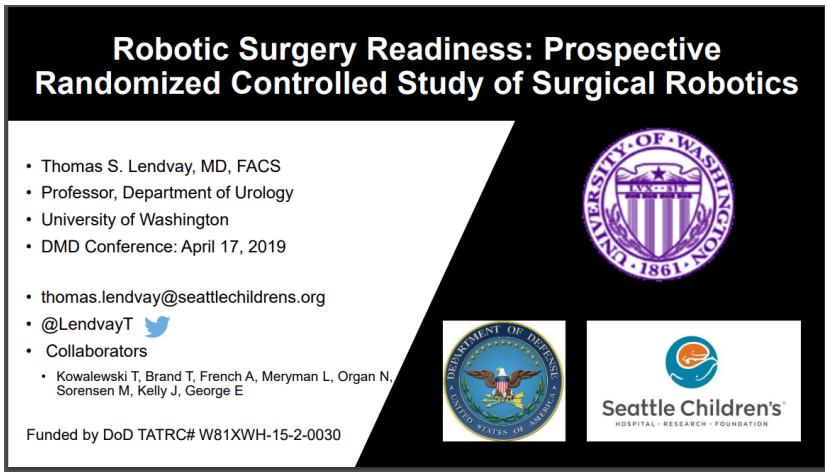


Figure 4. Task Time for subjects by group. Modules included in each group.

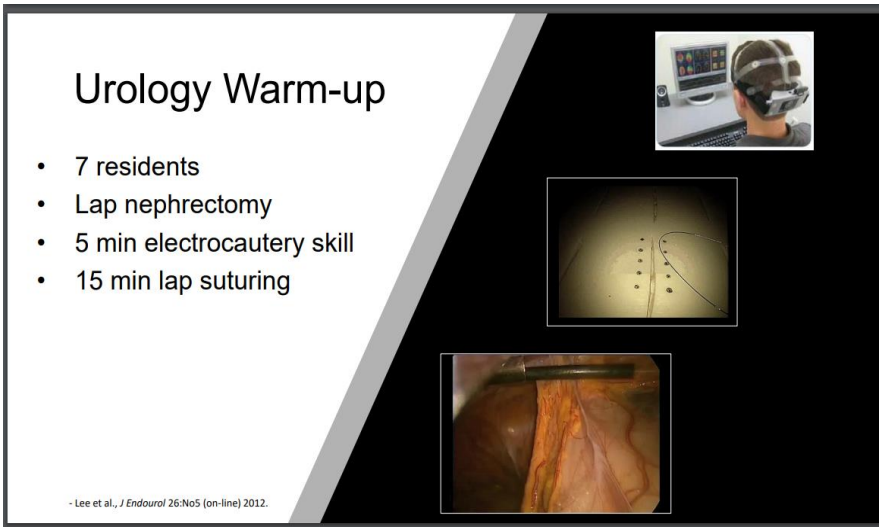
Appendix K

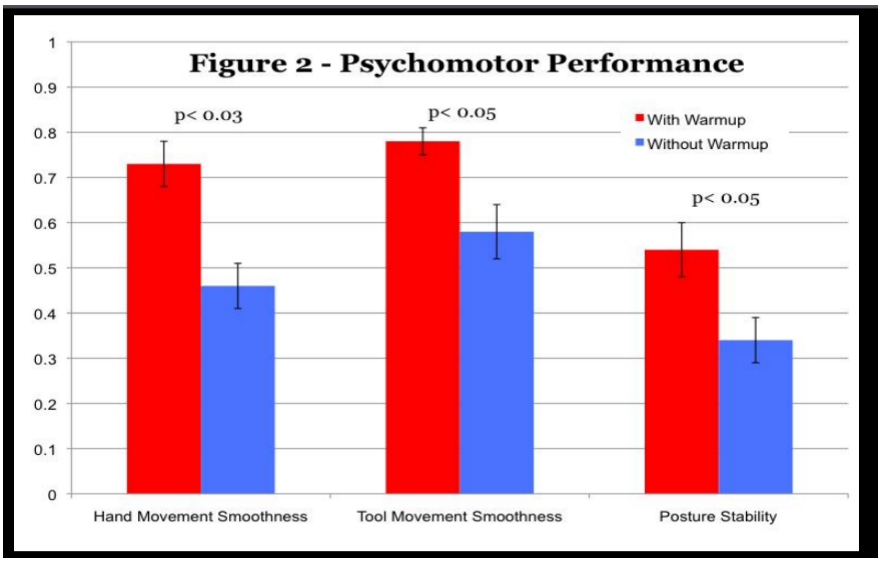
Aim 1 results were also presented at Design of Medical Devices Conference, University of Minnesota, April 16 -18, 2019

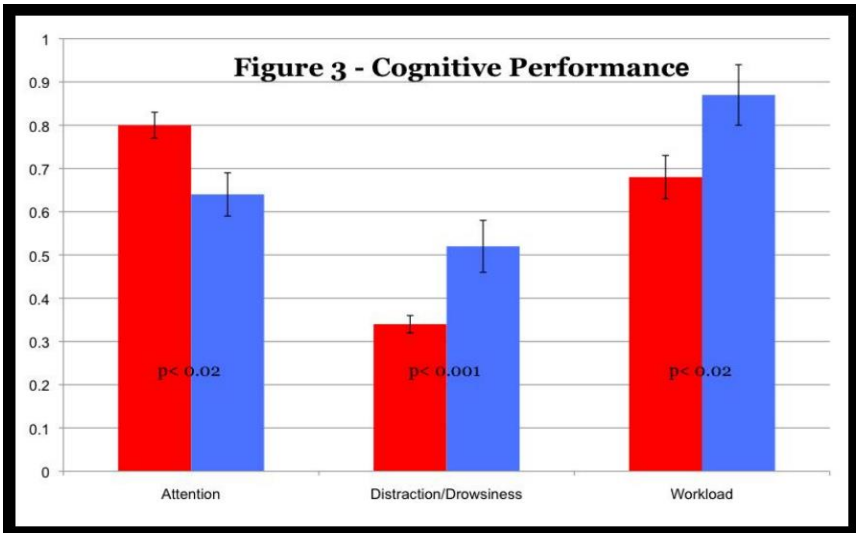

Robotic Surgery Readiness: Prospective Randomized Controlled Study of Surgical Robotics
• Thomas S. Lendvay, MD, FACS
• Professor, Department of Urology
• University of Washington
• DMD Conference: April 17, 2019
• thomas.lendvay@seattlechildrens.org
• @LendvayT
• Collaborators
• Kowalewski T, Brand T, French A, Meryman L, Organ N, Sorensen M, Kelly J, George E
Funded by DoD TATRC# W81XWH-15-2-0030
Seattle Children's
HOSPITAL · RESEARCH · FOUNDATION

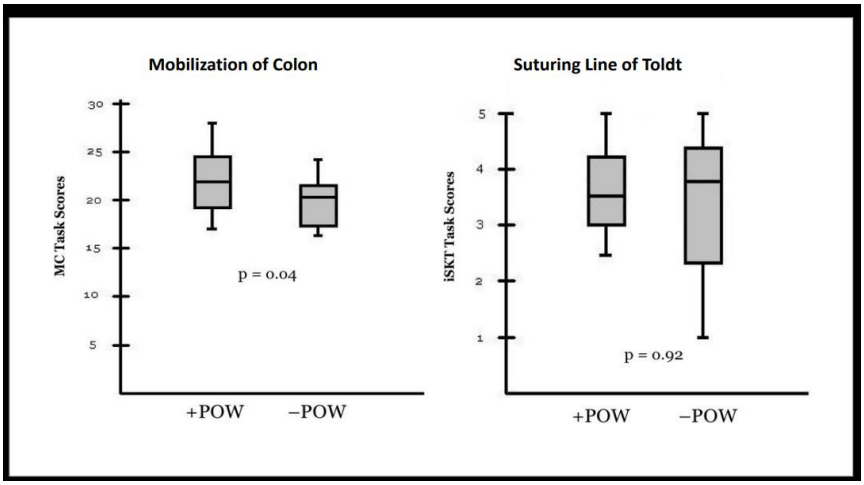

Decay in surgeon skill leads to ERRORS


Background
• 1970's behavioral psychology studies
• Warm-up decrement
• Decrease in performance after interval of rest
• Do et al. 2006
• Lap box trainer
• Bean transfer
• 25% improvement
Do et al. JSLS 10:297, 2006


Urology Warm-up
• 7 residents
• Lap nephrectomy
• 5 min electrocautery skill
• 15 min lap suturing
- Lee et al., J Endourol 26:No5 (on-line) 2012.


Figure 2 - Psychomotor Performance
p< 0.03
p< 0.05
p< 0.05
With Warmup
Without Warmup
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Hand Movement Smoothness
Tool Movement Smoothness
Posture Stability


Figure 3 - Cognitive Performance
p< 0.02
p< 0.001
p< 0.02
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Attention
Distraction/Drowsiness
Workload



Background

- 241 studies reviewed
- 6 were RCTs in humans
- All in laparoscopy
- All in trainees
- 196 surgeries
- Warm-up improved performance ($p<0.05$), 5/6 studies

- What about robot-assisted surgery?


Surg Endosc (2015) 29:1259–1269
DOI 10.1007/s00464-014-3811-4

REVIEW

The effect of warm-up on surgical performance: a systematic review

Gamal Abdalla · Erin Moran-Atkin · Grace Chen · Michael A. Schweitzer · Thomas H. Magnuson · Kimberley E. Steele

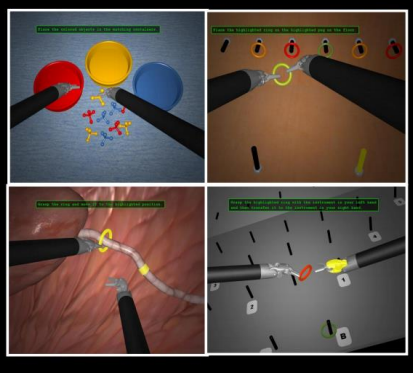
Hopkins

Seattle Children's
HOSPITAL • RESEARCH • FOUNDATION

Virtual Reality Robotic Surgery Warm-Up Improves Task Performance in a Dry Laboratory Environment: A Prospective Randomized Controlled Study

Thomas S Lendvay, MD, FACS, Timothy C Brand, MD, FACS, Lee White, BSc(Hons), PhC, Timothy Kowalewski, PhD, Saikiran Jonnadula, MD, Laina D Mercer, MS, Derek Khorsand, BSc(Hons), Justin Andros, BSc(Hons), Blake Hannaford, PhD, Richard M Satava, MD, FACS

- Hypotheses
  - Improves performance
  - Improves performance irrespective of skill level
  - Generalizes
- 50 faculty and residents
  - 3 trial sessions

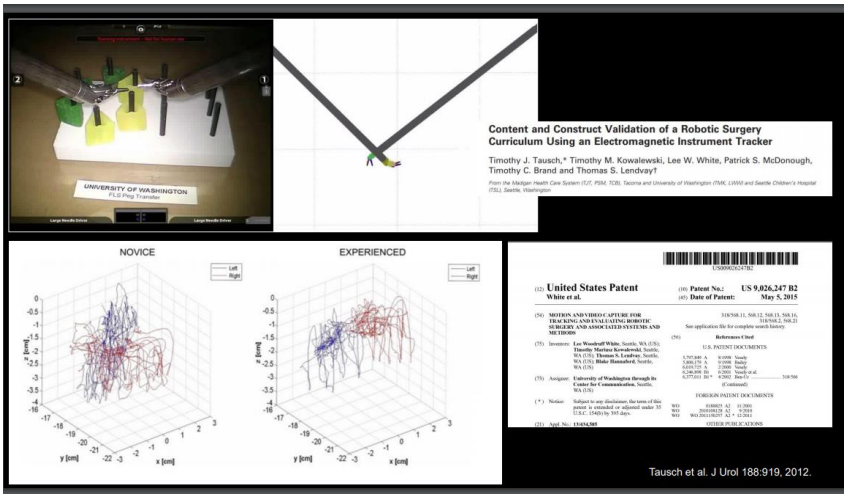


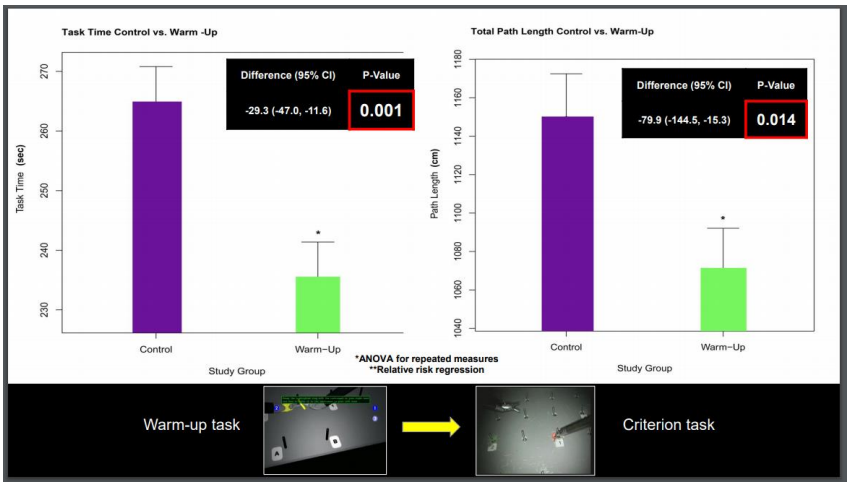
Lendvay et al. *Journal of American College of Surgeons* 216:1181-1192, 2013.

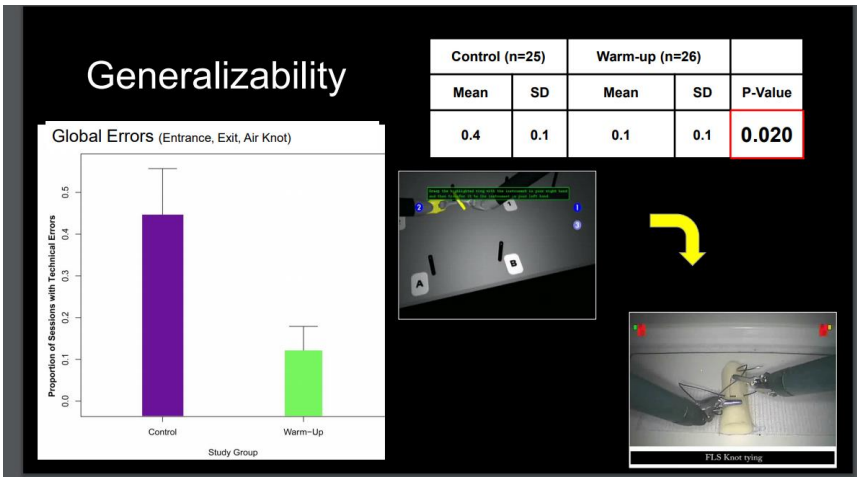
Funded by DoD TATRC # W81XWH-09-1-0714

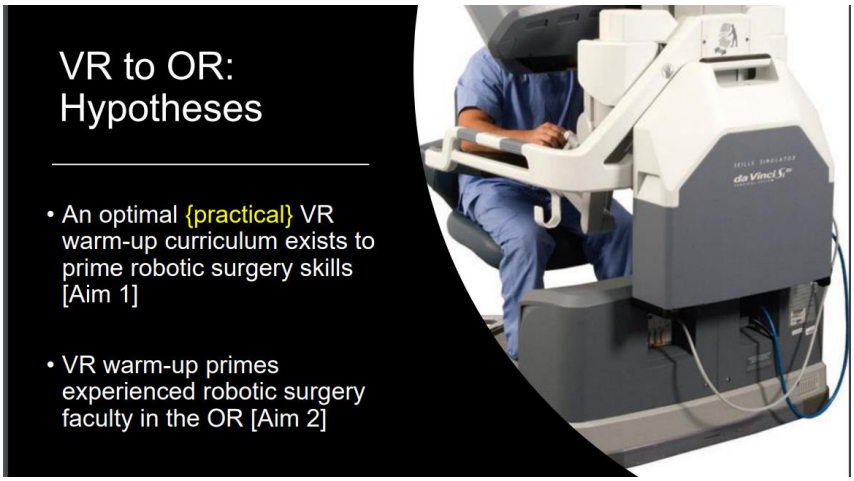

Military Significance
• Deployed surgeons
• Long periods of inactivity
• Returning surgeons
• Reintegration




UNIVERSITY OF WASHINGTON
FLS Peg Transfer
Content and Construct Validation of a Robotic Surgery Curriculum Using an Electromagnetic Instrument Tracker
Timothy J. Tausch,* Timothy M. Kowalewski, Lee W. White, Patrick S. McDonough, Timothy C. Brand and Thomas S. Lendvay†
NOVICE
EXPERIENCED
United States Patent
Tausch et al. J Urol 188:919, 2012.

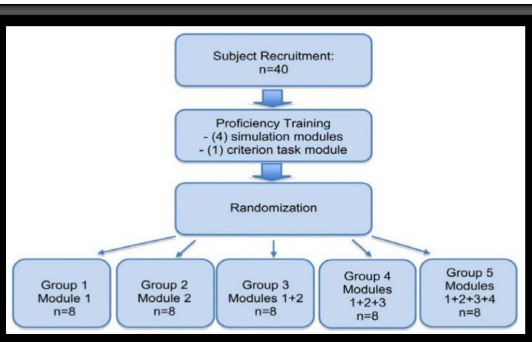

Task Time Control vs. Warm -Up
Difference (95% CI)
P-Value
-29.3 (-47.0, -11.6)
0.001
Task Time (sec)
Control
Warm−Up
Study Group
Total Path Length Control vs. Warm-Up
Difference (95% CI)
P-Value
-79.9 (-144.5, -15.3)
0.014
Path Length (cm)
Control
Warm−Up
Study Group
*ANOVA for repeated measures
**Relative risk regression
Warm-up task
Criterion task

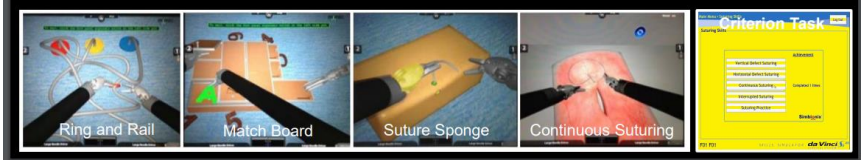

Generalizability
Control (n=25)
Warm-up (n=26)
Mean
SD
Mean
SD
P-Value
0.4
0.1
0.1
0.1
0.020
Global Errors (Entrance, Exit, Air Knot)
Proportion of Sessions with Technical Errors
Control
Warm−Up
Study Group
FLS Knot tying

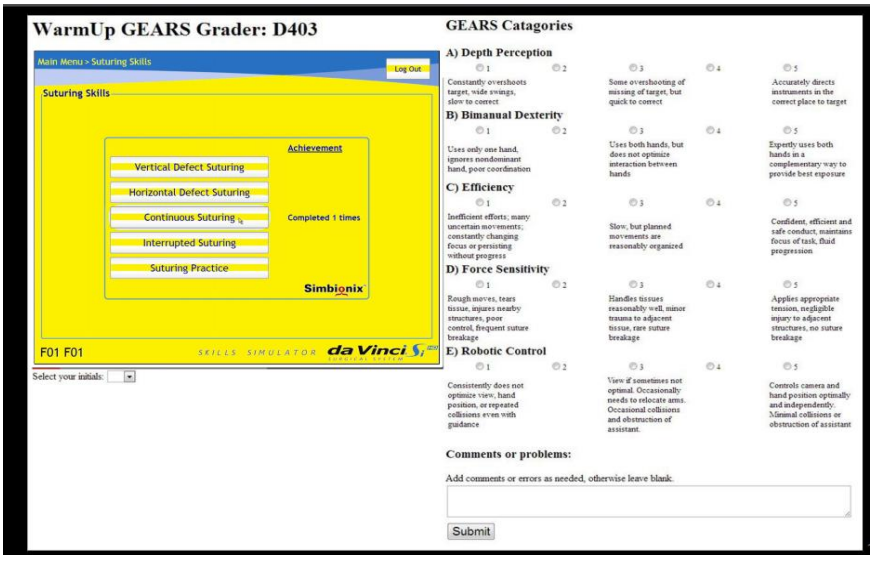

VR to OR:
Hypotheses
• An optimal {practical} VR warm-up curriculum exists to prime robotic surgery skills [Aim 1]
• VR warm-up primes experienced robotic surgery faculty in the OR [Aim 2]

AIM 1: Optimal VR Warm-Up

Methods







Methods and Demographics

- Five sites – private, academic, military
- 44 subjects randomized
- 6 trial sessions (practice effect)
- Track performance metrics
- Video assessed by experts
- Global Evaluative Assessment of Robotic Skills (GEARS)*

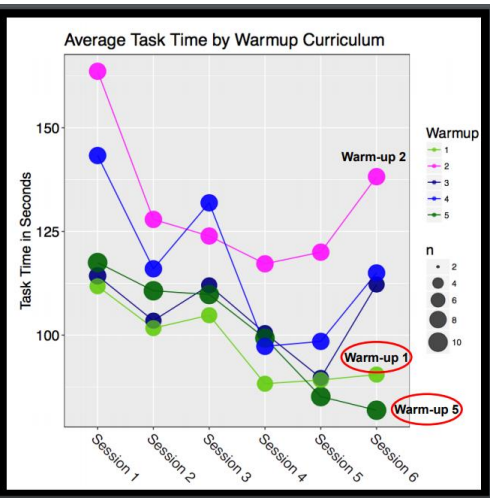
Gender	Female 31%	Male 69%
Experience	Faculty 30%	Trainees 70%
Disciplines	Urology 36%	GYN, Gen Surg, ENT 64%
Handedness	Right 85%	Left/Ambi 15%

*Goh et al. *J Urol*, 187:247-52, 2012.

Results: Task Times

- 1 and 5 - fastest
- 1 > 2
  - 31.3 (+/- 14.0) sec faster, p = 0.03
- 1 not different from 5
  - p = 0.70

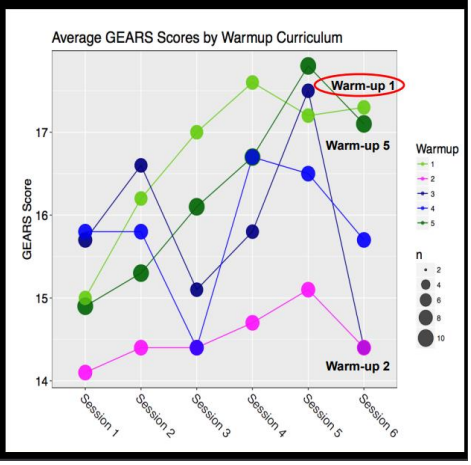
	1	2	3	4	5
Running Suture					
Ring and Rail					
Match Board					
Suture Sponge					



Results: Objective Skills

- 1 and 5 – best scores
- 1 > 2
  - 2.1 points higher (+/- 0.864), p = 0.02
- 1 not different from 5
  - p = 0.47

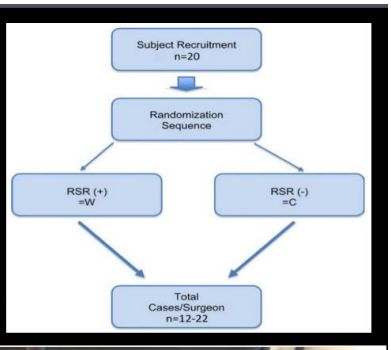
	1	2	3	4	5
Running Suture					
Ring and Rail					
Match Board					
Suture Sponge					



AIM 2: Warm-up Primes Surgeons

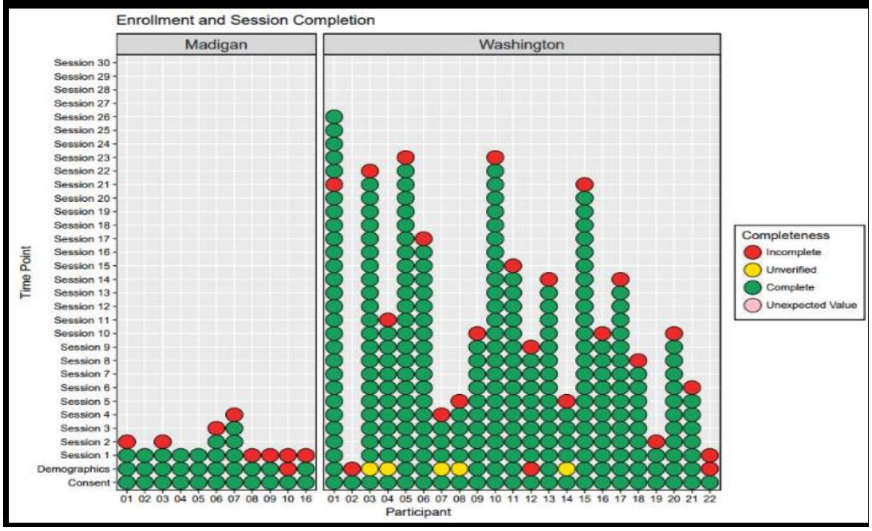
Methods

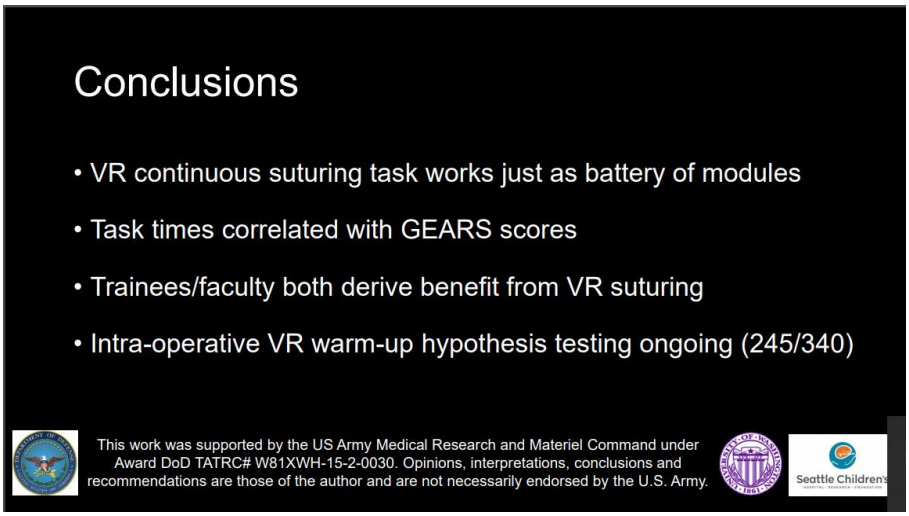
- Performance metrics (dVLogger)
- Video
- Assess with expert reviews
- 1st 15 minutes operative time
  - Warm-up v. no warm-up











Appendix L

Kelly, Jason "Analysis and Control of Temporal Biases in Surgical Skill Evaluation" PhD diss., University of Minnesota, 2020.

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF THE UNIVERSITY OF MINNESOTA

BY

Jason David Kelly

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Timothy M. Kowalewski, PhD

May, 2020

© Jason David Kelly 2020
ALL RIGHTS RESERVED

Acknowledgements

I would like to first thank my advisor Tim Kowalewksi for his guidance throughout my time in his lab. His understanding, support, and constructive criticism have provided an immense help in completing this project. Looking back I cannot imagine completing my graduate work with any other faculty member, and I hope to work as well together in the future.

I am grateful to all the members of the Robotic Surgery Readiness study at the University of Washington and University of Minnesota including: Thomas Lendvay, Lois Meryman, Anna French, and

Nicholas Heller, who were instrumental in achieving certain milestones and collecting data presented here.

Thanks to all who have been involved in the Medical Robotics and Devices Labo- ratory while I have been in it: Gillian McDonald, Chaitanya Awasthi, Mark Gotthelf, Mihai Duduta, Yusra Farhat Ullah, Bradley Drahos, Matthew Kubala, Ben Hamlen, Bin Fu, Mark Gilbertson, Trevor Stephens, Rebecca Smith, and several others. Whether we have worked together or not, the sense of community in our lab has made being a graduate student feel a little less like work.

I am additionally forever grateful for both of my parents, Jill and Andy, whom have always been there when I have needed them and raised me to be able to achieve new levels of success in our family. Lastly I would like to thank my wife, Katherine, who has been by my side throughout this process, and weathered all of the late nights, deadlines, failures and successes with me, helping make the journey more bearable.

Seattle Children's — Internal

Abstract

Objectively and accurately assessing the technical skill of a surgeon is critically important. The current gold standard relies on a panel of expert surgeons evaluating surgical video footage using structured survey instruments [1]. This is a prohibitively time-consuming process, thus leaving the majority of procedures unevaluated. Previous methods of evaluating skill remain prone to bias towards a surgeons' speed or task times, fueling the need to investigate the mech- anisms underlying human motion in favor of techniques impervious to biases.

The research objective of this work is to investigate the effects of time and speed on the relative accuracy of both human and computational methods of measuring surgical technical skill. Human methods consist of both expert and non-expert raters (faculty surgeons and Ama- zon Mechanical Turk crowd workers respectively). Computational methods consist of both neurophysiologically-derived measures from other disciplines and recent model-free machine learning methods.

This research objective is pursued by the following four specific aims:

Specific Aim 1: *Determine whether surgical motion segments are directly correlated to tangential velocity prediction models, and if the result is impervious to surgeon speed.*

The objective was to test the null hypothesis that there is no relationship between the minimum jerk trajectory velocity prediction model and increases in technical skills proficiency. Prior work in human reaching suggests that adherence to the minimum jerk model should increase as technical skill increases, for proficiency of reaching motions in stroke rehabilitation. This thesis investigates whether this phenomenon holds true in surgical technical skill during simulated dry lab tasks.

Specific Aim 2: *Implement a classification algorithm which uses recorded data to classify surgical skill in a manner which is impervious to task time.*

It was hypothesized that recent machine learning algorithms which exploit temporal du- ration, used on kinematic data from dry lab laparoscopic training tasks, would increase the performance of the current state-of-the-art computational methods of classifying surgical skill.

Specific Aim 3: *Determine how human ratings of surgical tasks are affected by video playback speed and duration.*

It was hypothesized that the perceived skill of a surgeon followed a unimodal function, in which human raters would experience an increase in perceived surgical skill as the speed of a surgical task video approaches the function's maximum, then immediately decreasing once being aware of the video's playback manipulation.

Specific Aim 4: *Measure the effect that pre-operative warm-up using validated virtual- reality simulator tasks has on practicing surgeons in real robotic surgeries as measured by the most accurate, least-biased methods detailed in the previous specific aims and prior art.*

This tested the hypothesis that pre-operative warm-up results in a measurable improvement in surgical technical skill among practicing surgeons (no novices) using surgical robots from live patients.

Seattle Children's — Internal

This research concluded firstly that neurophsyiologically-derived models of skill, specifically the minimum jerk model, do not necessarily extend to surgical settings. Surprisingly this research found the opposite, that surgeon experts exhibit movements which deviate further from the minimum jerk model. Second, a classification algorithm was created, using a bidirectional long short-term memory network which controls for task time, and is capable of classifying experts and novices with over 95% accuracy for tasks most resembling real surgery. This research brought about questions of label noise and accuracy, and emphasizes the importance of properly labeled data for machine learning algorithms. It was found that humans appear to have a speed bias in rating surgeons both for laparoscopic surgical training tasks as well as real robotic surgery procedures. Unexpectedly, this effect was more substantial for more expert performances and negligible for novice performers. Counter to our original hypothesis, expectations derived from biological motion models – that skill discrimination capability would unimodally decrease when video playback was obviously artificially sped up – were not met. Observer ability to discriminate skill continues well after people are cognizant of a video being played at quicker speeds and no discernible difference between biological-motion-relevant question groups (e.g. motion fluidity) and other questions appeared. Finally, a new dataset of robotic surgeries was introduced, with 343 videos of robotic surgeries including tooltip kinematic data. Evidence obtained from motion metrics, crowd ratings, and faculty surgeon ratings suggest that no measurable warm-up effect was present in our population of 41 practicing surgeons; no evidence supported the use of warm- up.

Seattle Children's — Internal

Contents

Acknowledgements **i**

Abstract **i**
i

List of Tables **i**
x

List of Figures **x**
i

1 Introduction **1**

2 Background **4**

2.1 Project Description 5

2.1.1 Motivation and Objectives 5

2.1.2 Crowd Sourcing 6

2.1.2.1 Likert Scale Scoring 8

2.1.3 Hand Movement Research 1
2

2.1.3.1 Measures of Movement Smoothness 1
2

2.1.3.1.1 Submovements 1
2

2.1.3.1.2 Spectral Arc Length 1
3

2.1.3.1.3 Minimum Jerk 1
6

2.1.3.2 Objective 1
8

2.1.4 Computational Skill Evaluation 1
8

2.1.4.1 Neural Networks 1
8

2.1.4.2 Recurrent Neural Networks 2
1

2.1.4.3 Long Short-Term Memory 2
1

2.1.4.4 Automated Performance Metrics 23
2.1.4.5 Video Motion Data Analysis 23
2.1.4.6 Lack of Quality Data 26
2.1.4.7 Objective 28
2.1.5 Biological Motion Perception 28
2.1.5.1 Visual Stimuli Research 29
2.1.5.2 Response to Abnormal Dynamics 30
2.1.5.3 Objective 30

Seattle Children's — Internal

.1.6 Datasets to be used in this work 32
6.1 Basic Laparoscopic Urologic Study 32
6.2 Robotic Surgery Readiness Study 32
.1.7 Linear Mixed Effects Modeling 33

3 Neurophysiological Models 36

3.1 Testing the Relationship of Minimum Jerk Model to Surgical Skill 36
.1.1 Methods 36
.1.2 Results 37

4 Computational Skill Evaluation 44

4.1 Bidirectional Long Short-Term Memory for Surgical Skill Classification of Temporally Segmented Tasks [69] 45
4.1.1 Abstract 45
4.1.2 Introduction 46
4.1.3 Methods 47
4.1.3.1 Dataset 47
4.1.3.2 Skill Classification from Temporal Segmentation 49
4.1.3.3 Data Partitioning 49
4.1.3.4 Window Parameter Selection 49
4.1.3.5 LSTM Parameters and Architecture 50
4.1.3.6 Crowd Reassessment 51
4.1.4 Results 52
4.1.4.1 Skill Classification from Temporal Segmentation 52
4.1.4.2 Crowd Reassessment 57
4.1.5 Conclusion 60
4.2 Classification Using Video Data 62

5 Speed Perception 65

5.1 The Effect of Video Playback Speed on Surgeon Technical Skill Percep-

tion [73] 66
.1.1 Abstract 66
.1.2 Introduction 67
2.1 Biological Motion Perception 68
2.2 Objective Metrics 68
.1.3 Methods 71
3.1 Dataset 71
3.2 Experiment 1: Technical Skill Perception 72
3.3 Experiment 2: Speed Perception 73
3.4 APM Validation at Each Playback Speed 73
.1.4 Results 74
4.1 Technical Skill Perception 74
4.2 Playback Speed Perception 76
.1.5 Conclusion 77
5.2 The Effect of Video Playback Speed on Perception of Technical Skill in Robotic Surgery [81] 83
.2.1 Abstract 83
.2.2 Introduction 84
2.1 Changes in Perception due t o Speed 86
2.2 Technical Skill Ranges 87
.2.3 Methods 88
3.1 Dataset 88
3.2 Technical Skills Perception at Different Playback Speeds 89
3.3 Sub-Task Level Skill Labeling 90
.2.4 Results 90
.2.5 Data Demographics 90
5.1 Technical Skill Perception 92
5.2 Sub-Task Level Labeling 93
.2.6 Conclusion 94

6 Robotic Surgery Readiness Study 96

6.1 Robotic Surgery Readiness: A Randomized Controlled Study of Virtual Reality Warm-Up Prior to Robot-Assisted Surgery [93] 97
.1.1 Abstract 97
.1.2 Introduction 98
.1.3 Methods 101
6.1.3.1 Aim 1 101
1.1 Subject Recruitment 101

1.2 dVSS Modules 102
1.3 Proficiency Testing 103
1.4 Video Review 104
1.5 Statistical Models 105
6.1.3.2 Aim 2 106
2.1 Subject Recruitment 106
2.2 Randomization and Warm-Up Protocol 106
2.3 OR Data Capture 106
2.4 Data Aggregation and Manipulation 106
2.5 Video Segmentation 107
2.6 Objective Metric Calculation 107
2.7 Video Review 108
1.4 Results 108
6.1.4.1 Aim 1 108
6.1.4.2 Aim 2 112
2.1 GEARS Review Outcome 112
2.2 Objective Metric Outcomes 114
1.5 Discussion 116
1.6 Conclusions 117
6.2 Temporal Variability in Surgical Technical Skill Evaluation [117] 118
2.1 Abstract 118
2.2 Introduction 119
2.3 Methods 121
3.1 Dataset 121
3.2 Crowd Evaluation 122
3.3 Statistical Analysis 123
2.4 Results 124
2.5 Conclusion 126

7 Conclusion and Discussion 128

7.1 Contributions 129
7.2 Limitations 131
0.1 Speed Perception Studies 131
0.2 Long Short-Term Memory Study 131
0.3 Robotic Surgery Readiness Study 131
0.4 Rating Variability Study 131
7.3 Future of Surgical Skill Research 132

References **133**

Appendix A. Robotic Surgery Readiness Supplemental Figures **149**

.0.1 Kinematic Metrics 153

.0.2 Event Metrics 153

Seattle Children's — Internal

List of Tables

2.1 Likert-scale technical skill perception questionnaire for video-based eval uation of a surgical task, from four domains of the GOALS assessment metric [24]. 10

2.2 Likert-scale technical skill perception questionnaire, from the five of the 6 domains in the GEARS assessment metric. Autonomy is left out, as no audio or video of the operating room is provided i n our datasets. 11

2.3 Summary of popular skill evaluation studies, and lack of high-volume data. 27

3.1 ANOVA test results for the mixed effects model on the minimum jerk viability as a predictor variable for task score value. 40

4.1 Main hyperparameters tested during model evaluation in Section 3.1, with most optimal results i n bold. 53

4.2 Accuracy results computed by the bidirectional LSTM for each BLUS task. 55

4.3 Correlation coefficients for the relationship between GOALS scores and predicted scores from the LSTM for every performance from each of the four BLUS tasks, including intermediate level performers, from Section 3.1. 55

5.1 Likert-scale speed perception questionnaire for a manually sped-up video. 68

5.2 Results from the linear mixed effects model testing speed and time spent on evaluation. 90

5.3 The types of robotic surgery performed and demographics of the surgeons from the 56 videos used. 91

6.1 The five groups of warm-up modules used for aim 1. Each group consisted of different combinations of warm-up modules shown to help elevate skill. 105

Results from Friedman test with kinematic and event outcomes. 115

Correlation values for each of the metrics measured with GEARS scores. 115

Summary information for 12 videos used from the RSR study 122

A.1 Recorded event measures for the RSR study 153

List of Figures

2.1 Relationship between speed of task completion and crowd assessed score. Surgeons with a longer task time on average receive lower scores, as rated by crowds. 6

2.2 Original normal curve found through Galton's crowd-sourcing experi- ments with ox weights [17]. The mean of these data converge to the correct weight of the ox, within one pound. 7

2.3 Previous research found a high correlation between faculty surgeon review and non-expert crowd review of surgical technical skill [21] 8

2.4 Typical movements(bold lines) and their extracted submovements(fine lines) from first and last days of a therapy session for a stroke patient [11]. 12

2.5 The effect of the SAL metric on varying lengths of movements [28]. Longer movements provide a lower SAL measurement. 15

2.6 The effect of temporal movement scaling on SAL vs. SPARC [28]. SPARC is relatively impervious to movement scaling. 16

2.7 Predicted(solid lines) and measured(dashed lines) hand trajectories over- laid [3]. The minimum jerk model provides accurate predictions for the movements of recovering stroke patients. 17

2.8 An example of the architecture of a neural network [34]. Inputs are sent through layers in which activation functions and array manipulations are used to arrive at a single output signal. 20

2.9 An example of the architecture of a long short-term memory network. A series of gates are used to create a sense of context for past mistakes and successes. 22

2.10 The detected SIFT points from a dry lab laparoscopic surgical task image (computed using MATLAB computer vision toolbox). 24

2.11 Comparison of corner detection boundaries for Harris and Shi-Tomasi detection algorithms. Shi-Tomasi has a more rigid set of constraints, and will detect more corners because of these boundaries. 25

2.12 Top: Discriminated motion counts corresponding to each class of STIPs. Bottom: Frame kernel matrices example, displaying the relationship be- tween various segments of a video [55]. 27

2.13 Example of a point light walker. Each of the dots represent a limb or joint of the human body 29

2.14 Illustration of a PLW representing an animal (right) and the change in perceived size as the gait increases (left). This suggests a link between speed and biological motion perception [66]. 30

2.15 Frames from a (left to right) suturing, cutting, and peg transfer task in the BLUS dataset. 32

2.16 Range of crowd-assessed surgical skill evaluation mean scores for each of the BLUS tasks. On average clipping tasks receive higher scores than, for example, suturing tasks. 33

2.17 Frame from a video in the RSR data set. Typical robotic surgeries recorded were prostatectomy, hysterectomy, and partial nephrectomy, with all procedures being recorded in Washington state. 34

3.1 All Peg Transfer task segments overlaid to show the general shape of peg transfer task grasp segments. One single line represents one single move- ment by a surgeon. Movements were defined as starting and ending when the hand reached nonzero speed. Time is normalized by the duration of each segment and velocity is scaled by the peak velocity magnitude of that segment. 38

3.2 Example of segmented grasps, from the first performance of the peg trans- fer tasks only. Showing the data in this manner allows seeing the vari- ability that exists throughout the different movements in a surgical task performance. 39

3.3 Unimodal movements from peg transfer tasks fitted to minimum jerk movements described in Chapter 2.1.3.1.3. 40

3.4 Fixed effects model of observed vs. predicted skill, for all BLUS tasks combined, using minimum jerk model, subject and task identity, in which the fixed effects accounts for a constant intercept for all of these factors. The fixed effects model behaves quite poorly 41

3.5 Linear Mixed effects model's predicted vs. observed skill using minimum jerk model fit error, task, and subject indentity, in which this model accounts for varying intercept values for each task, and subject. This model is able to predict outcomes much better than the fixed effects model, with a correlation between actual and predicted skill of 0.826. 42

3.6 LME predictions grouped by task, with colorbar illustrating total time for task completion. There is a strong correlation between observed task scores and predicted scores, with slower performances usually receiving poorer of both. 43

4.1 Example left hand tool acceleration time series with the sliding time window extraction method applied to both a novice series and an expert series. A novice performance is indexed with a window overlap of 105 time indices, and an expert performance is indexed with a window overlap of 50 time indices. 51

4.2 Process of computing final suturing performance prediction. All ten can didate predictions of a performance are averaged at each index to compute the final prediction. 52

4.3 The LSTM's prediction for all performances from each of the tasks in the BLUS dataset. The LSTM was only trained on experts and novices (summary score labels with cross validation), from Section 3.1. 54

4.4 The LSTM predictions trained on crowd scores, for performances of sutur- ing tasks as rated by faculty surgeons, which has a Spearman correlation

of 0.89, from Section 3.1. 56
4.5 The perceived score of the LSTM on the left, with a video on the right of the same task. 57
4.6 Reassessment of misclassified suturing and peg transfer performances sug- gest crowds agree more with LSTM than previous crowd ratings, from Section 3.2. 59
4.7 A bar plot of the magnitude of attenuation of the crowds to agree more with the LSTM on score reassessment, than with the original scores obtained. 60
4.8 The frames of a novice video with the amount of points belonging to each motion class being shown. 62
4.9 The frames of an expert video with the amount of points belonging to each motion class being shown. 63
4.10 Frame kernel matrix for a portion of a time series, to visualize similarity between frames. 63
5.1 All mean crowd evaluations from each novice and expert peg transfer task at various video playback speeds. (Each solid marker indicates N = 40). 74
5.2 The efficiency subdomain as compared to the average of the other GOALS subdomains, for experts and novices. 75
5.3 Objective technical skill metrics for the obvious experts and novices across various speeds. 76
5.4 Difference in mean between skill levels from the highest performing APM and from crowd scores, with a 95% CI in the shaded regions. 77
5.5 Comparison of Novice and Expert Speed Perception at each playback speed. (There were $N = 40$ *unique human evaluations/skill level/playback speed*. The error bars represent 95% confidence intervals.) 78
5.6 Speed Perception compared with technical skill perceived at various speeds. 79
5.7 Crowd GOALS evaluations for Experts and Novices at various playback speeds. Each point is $N = 40$, with 3600 total evaluations. 81
5.8 Crowd GOALS evaluations for Experts and Novices at various playback speeds. Each point is $N = 40$, with 3600 total evaluations. 82
5.9 Frame from a video collected for this study, recorded from a daVinci surgical robot, Intuitive (Sunnyvale, CA). 85

Seattle Children's — Internal

5.10 Illustration of research results (left) which found the frequency of gait in simulated animals (right) affects the estimated size [**?**]. 86

5.11 All mean crowd evaluations (bold lines) from each proficient and expert surgeon at various video playback speeds. Each single surgeons video (semi-transparent colors) indicates ratings from N = 40 turkers. 91

5.12 The efficiency subdomain as compared to the mean of the other GEARS subdomains, for expert and proficient surgeons. 92

5.13 GEARS scores given to the entire 15 minute video of a performance, compared to only the first minute of the same video. 93

6.1 The popularity in robotic surgery has experienced a significant increase in recent years..... 99

6.2 The da Vinci Skill Simulator (dVSS) used for surgical simulation modules, created by Mimic Technologies..... 100

6.3 The four dVSS warm-up modules used for aim 1 of the RSR study in- cluding Ring and Rail 2, Match Board 3, Suture Sponge 3, and Running Suture (left to right). 103

6.4 The randomization setup for the 5 warm-up groups in aim 1 of the RSR study 104

6.5 Enrollment demographics for participants in aim 1 of the RSR study 109

6.6 The task time (TT) for the five different warm-up module groups, through- out the six completed sessions.....110

6.7 The GEARS scores for the five different warm-up module groups, through- out the six completed sessions.....111

6.8 Enrollment demographics for participants in aim 2 of the RSR study 112

6.9 The correlation of the GEARS scores from faculty surgeons and C-SATS scores was weak, with a value of 0.1714, 95% CI from -0.1287 to 0.4426. 113

6.10 The effect of Si and Xi da Vinci platforms for surgeons who performed and did not perform a warm-up module before surgery. No significant difference exists between these four groups..... 114

6.11 The effect of performing a warm-up module for each surgeon in the RSR study. With a clear normal distribution, no significant effect was measured. 116

6.12 The attention question inserted to the GEARS questionnaire, for quality assurance purposes. Note that skimming the instructions likely results in incorrect answers. 123

6.13 All mean crowd evaluations from each proficient (red) and expert (blue) surgeon at each minute of the performance. Surgeons previously ran- domized to the control (solid lines) group normally started surgery with- out any intervention or pre-operative warm-up used. Warm-up (dashed lines) group surgeons reviewed a virtual reality warm-up module prior to surgery. (The warm-up hypothesis from the original randomized study is not being tested or evaluated in this research, only temporal variation in ratings is.....124

6.14 Comparison of previously-obtained scores for 15 minute video (95% CI) and average score for evaluation of 1 minute segment. (PW = Profi- cient/Warmup; PC = Proficient/Control; EW = Expert/Warmup; EC=Expert/Control) 125

6.15 Standard deviation from the LME model for random effects from each individual crowd rater, compared with the fixed effect of time. The rater effect is clearly larger, although time is still significant in predicting out- comes..... 126

A.1 Video of an Si daVinci surgery, after editing the video to the correct times

and placing black boxes over the identifying pieces of information. 150

A.2 3-D Plot of the kinematic tooltips from a robotic surgery which corre- sponds with video edited at the same time range. USM 1-3 refers to the Universal Serial Manipulators, the surgical arms. 150

A.3 A view of the website faculty surgeons would see when visiting to evaluate robotic surgeon technical skill. 151

A.4 Cumulative positions of several combined tasks from the RSR study, in both 3-D coordinates, as well as a histogram for frequency of position in each axis. 152

Seattle Children's — Internal

Chapter 1

Introduction

This thesis aims to discuss previous and current work in computational surgery, and more specifically, surgical skill evaluation. Skill evaluation is a topic becoming popular in several research labs, but it is apparent that accurately evaluating technical skill is a multi-faceted problem, with several specific areas of research which should be investigated. Moreover, previous research appears to reward time of task completion, perhaps unfairly, to surgeons. The work presented in this thesis aims to both summarize previous work as well as test the existence of this bias in surgical settings and create techniques which control for these biases.

The research from this thesis begins by examining hand movements models which were derived from neurophysiology research involving stroke patients (Chapter 3). It was found that the form of hand movements of recovering stroke patients most closely agreed with the minimum jerk velocity profile. To test these models and control for the speed of performances, individual point-to-point movements from surgical tasks were extracted and normalized to use with statistical models. The results show a surgeon is surprisingly more likely to be novice-like if they exhibit movements which correspond to the minimum jerk model, contrary to what prior research would suggest.

Summary metrics of performances can be biased to give quicker surgeons better metric scores due to the way they are calculated. In Chapter 4 machine learning models were used in an attempt to classify the skill of a surgeon, in a way which does not bias task time. Multidimensional kinematic data was used to train a model, in which the data was segmented into small multi-second chunks. Training the model in this manner

Seattle Children's — Internal

allowed the model to learn aspects of movements without being biased to reward faster or slower surgeons. This was able to classify surgeons into skill levels with over 95% accuracy for suturing tasks which most closely resemble surgical procedures. However, reassessing misclassified surgeons led to questions about label noise, and reiterates the need for accurate labeling of surgical data.

The third aim was then addressed by examining the effect of video playback speed on surgical skill perception (Chapter 5). The goal was to learn whether humans perceive technical surgical skill differently when viewing video at quicker playback speeds, and if this continues once they are aware of the video being sped up. When working with laparoscopic simulation tasks it was found that skill was more easily discriminated when a video was manipulated to play faster. Additionally, this phenomenon extended past the point at which users were aware of the video being sped up, rejecting the null hypothesis that the perceived increase in skill was reliant on being unaware of the playback speed being changed. In an effort to create a more statistically powerful finding, this study was repeated for a larger dataset of robotic surgery procedures, which also found increased perceived skill levels at quicker playback speeds.

Finally, the fourth aim was addressed in Chapter 6, during which a new dataset of robotic surgery procedures was introduced. The results from this research, as measured by motion metrics, crowd scores, and faculty ratings, concluded that participating in a virtual reality warm-up module prior to robotic surgery did not significantly elevate the technical proficiency of the surgeons involved in this study. Nevertheless, this dataset will allow new and innovative techniques to be used in the future in pursuit of solving surgical skill evaluation challenges. This chapter also reviews work examining the temporal variability in skill evaluation, finding that for longer duration surgical procedures, summary metrics may hinder in properly conveying the overall skill of a surgeon. Measures which report the variability of surgical skill may be required.

This thesis will conclude with a chapter summarizing the main findings of this re- search, discussing limitations of the work presented, and discussing future work in the field of surgical skill evaluation (Chapter 7).

Seattle Children's — Internal

- ## Chapter 1: Introduction

- **Chapter 2: Background** - Briefly presents the history, related work, and need for improvement in the area of computational surgical skill evaluation.

- **Chapter 3: Neurophysiological Models** - Describes the process of evaluating the relationship between the minimum jerk model and surgical skill.

- **Chapter 4: Computational Surgical Skill Evaluation** - Provides an introduction to a machine learning algorithm which can accurately evaluate the technical skill of a surgeon.

- **Chapter 5: Speed Perception** - Provides an analysis and explanation for how people could be biased in evaluating surgical skills based on the perceived speed of a surgeon.

- **Chapter 6: Robotic Surgery Readiness Study** - Presents the need for the robotic surgery readiness study, explains the methods of gathering data, and the results of the study, in addition to studying how human perception of skill can vary in longer duration recordings.

- **Chapter 7: Discussion** - This chapter concludes the thesis work and provides guidance for how to proceed and improve upon this work in the future.

Seattle Children's — Internal

Chapter 2

Background

Large ‘crowd-sourced’ groups of non-experts are surprisingly accurate at evaluating surgical tasks [1], though it remains unclear as to ‘how’ or ‘why’. However, as shown in [2], human perception can be affected by various physical features such as size or speed. Due to this, work must be done to create an unbiased metric for viewing complex human motions using approaches including algorithms impervious to bias. Previous research has shown that optimized velocity trajectories of point-to-point movements can be predicted using methods such as minimizing the jerk of the tangential velocity of a movement [3]. However, surgeries involve very complex movements, and the mapping of these trajectories to a performer’s technical skill is not yet clear.

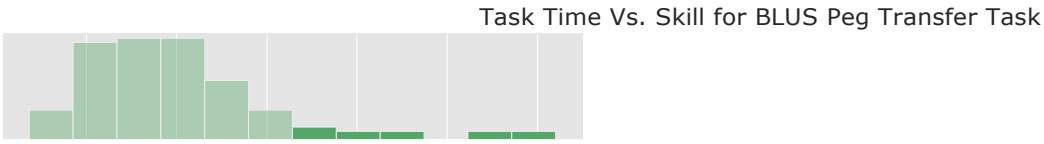
The objective of this research is to elucidate the mechanisms that underlay human perception of surgical skill, computational models, and velocity trajectory movements which correlate to surgeon technical performance, and accurately measure the perception of skill as a function of video playback speed.

Seattle Children's — Internal

1.1 Motivation and Objectives

The prevalence of surgical errors in the U.S. makes it essential to develop algorithms capable of learning how surgical skill can be classified [4], [5]. Part of this effort is learning what features of a procedure can be objectively determined to be expert-like. In the past, hidden markov models have been used to assess surgeon skill, but these are not incredibly accurate [6–8]. While they perform well with small amounts of data, they are not able to become increasingly accurate as more data is available. Compared to newer machine learning techniques, hidden markov models are relatively simple and will not be able to extract as complex information as other more recent techniques [9]. Much success in other academic areas has been the result of variants of recurrent neural networks, which can arm a model with a sense of memory to remember previous events. Using a model such as this with kinematic features from tooltips would be an immensely valuable way to classify surgical skill.

There is a plethora of previous research regarding skill, claiming certain tangential velocity hand movement models are related to skill of a performer, but much of this only relates to simple point-to-point movements [3, 10–12]. Surgery involves complex, but also repetitive movements. There are many factors at play which influence a surgeon's relative skill. The single metric which most closely correlates to the skill of a surgeon is using the time of task completion, or task time [13]. This relationship can be seen in Figure 2.1 for the BLUS peg transfer task. This is not perfect by any means though, and is just one of the factors at play when evaluating a surgeon's skill. There must be more in depth research done to evaluate this phenomenon, and learn whether phenomenal surgeons operating slowly can be perceived as experts or if crowds are rating fast surgeons well due to the quick movements they are seeing.



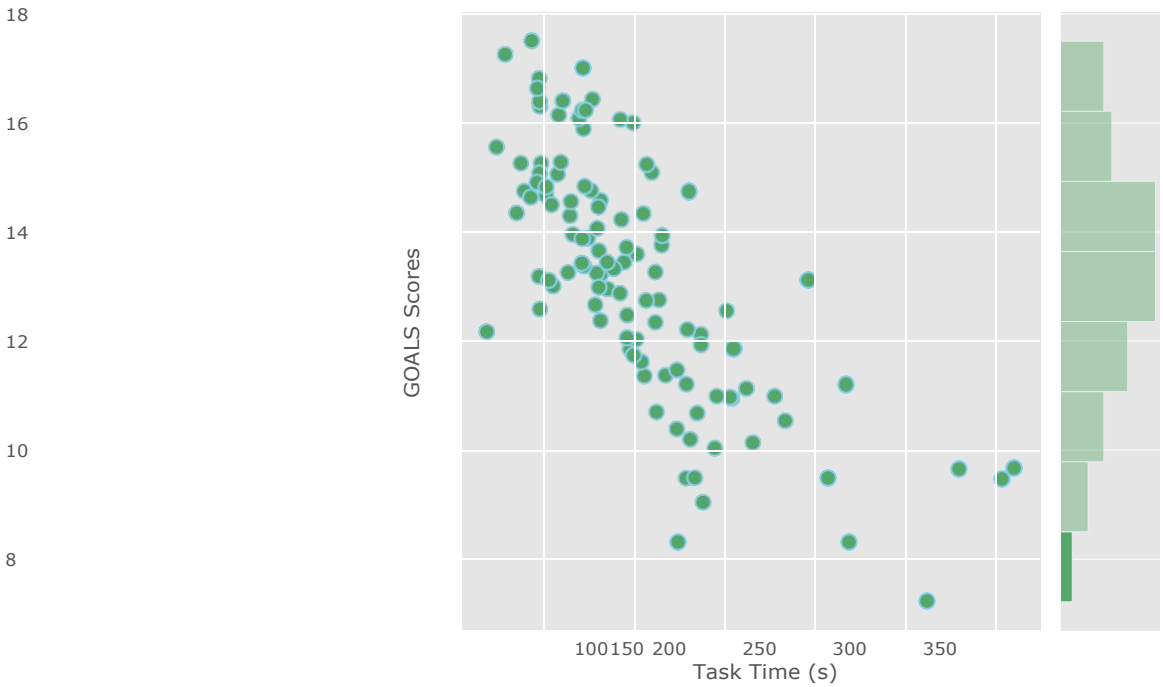
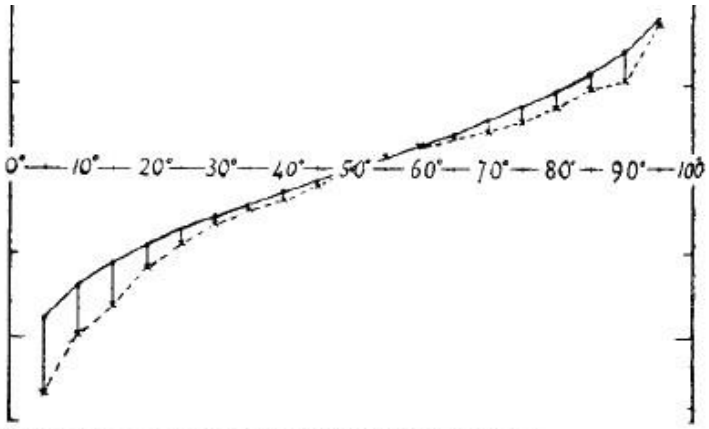


Figure 2.1: Relationship between speed of task completion and crowd assessed score. Surgeons with a longer task time on average receive lower scores, as rated by crowds.

1.2 Crowd Sourcing

Previous research has shown that when a large population of diverse people are asked certain questions, the truth emerges from the responses as the population size grows [14]. [15] defines crowdsourcing as, "*the act of outsourcing tasks originally performed inside an organization, or assigned externally in form of a business relationship, to an undefinable large, heterogeneous mass of potential actors*". Essentially, collective knowledge from a crowd can more easily reflect the true answer of specific types of questions, than can any one person. This was first discovered by statistician Sir Francis



The continuous line is the normal curve with p.e.=37.
The broken line is drawn from the observations.
The lines connecting them show the differences between the observed and the normal.

Figure 2.2: Original normal curve found through Galton's crowd-sourcing experiments with ox weights [17]. The mean of these data converge to the correct weight of the ox, within one pound.

Galton, the father of "eugenics". Galton famously came across a county fair competition in 1906, in which visitors were asked to guess the weight of an ox. He later used the responses to run some statistics. To his surprise, the mean of all the guesses resulted in 1197 pounds, the actual weight being 1198 pounds (Fig. 2.2) [16, 17]. This finding was later elaborated on, finding that the crowds needed to be both large and diverse, or this would not work accurately. There are debates about how many workers must be used to attain a satisfactorily 'true' answer, but generally, higher populations yield more accurate results [18].

These methods have also been used in the past to allow rating skill of a surgical task or study [19]. The current gold standard for rating surgeons is a faculty panel of surgeons which rate surgeries one at a time. This is very time-consuming and suboptimal, as faculty surgeons could spend their time with more valuable tasks. Crowdsourcing workers to rate surgical tasks has proved to be a valid method of evaluation with high correlation to the gold standard (Fig. 2.3) [20, 21].

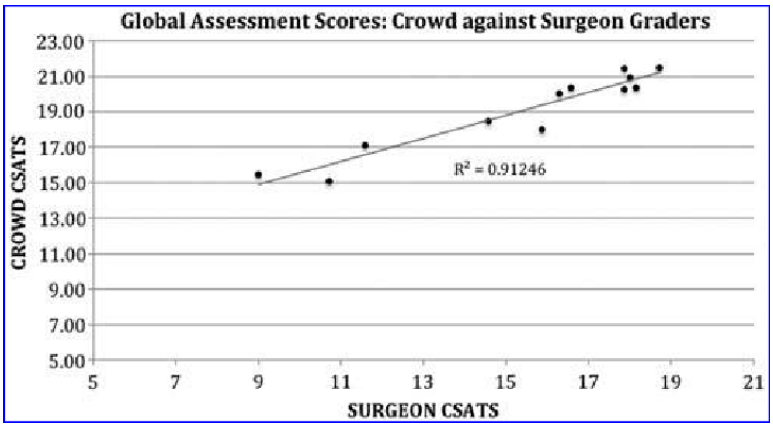


Figure 2.3: Previous research found a high correlation between faculty surgeon review and non-expert crowd review of surgical technical skill [21]

There are several platforms currently used for crowdsourcing tasks, such as Idea Bounty, Innocentive, CrowdSpring, and Amazon Mechanical Turk. Amazon's Mechanical Turk has become popular for businesses and companies to be able to request workers to complete "Human Intelligence Tasks", or HITs, for some amount of compensation. Very simple tasks such as completing surveys or watching videos, could be assigned. Once workers gain experience, and gain reliability, users can be offered more compensation to complete tasks, which filter out workers that may not put in a great amount of effort or focus. Mechanical Turk has been used previously in surgical crowdsourcing studies, finding that crowds give a 100% matching pass/fail rate of what faculty surgeons would rate when reviewing video of surgeons [1], [19, 22, 23].

2.1 Likert Scale Scoring

A Likert scale is a type of scale in which the user has a choice of several pre-defined options for how to answer. This could be a question in which the user answers with numbers from 1-10, or even answers such as "Likely", "Not at All Likely", "Very Likely", and so on. When both faculty surgeons and non-expert crowd workers evaluate surgical

tasks, they are most likely using a Likert scale to assess the surgeon. Two assessment methods will particularly be discussed several times throughout this research, the Global Evaluative Assessment of Robotic Skills (GEARS) and the Global Operative Assessment of Laparoscopic Skills (GOALS). These consist of five and four subdomain questions, respectively, which ask the rater to evaluate various aspects of the surgeon's skill, such as depth perception, tissue handling, or robotic control, on a scale of 1-5. Scores for each subdomain are totaled for a range of cumulative scores for GOALS being between 4-20 and for GEARS being 5-25, with higher scores inferring a more expert surgeon. Both scales are shown in Table 2.2 and Table 2.1.

Seattle Children's — Internal

Table 2.1: Likert-scale technical skill perception questionnaire for video-based evaluation of a surgical task, from four domains of the GOALS assessment metric [24].

Score	
	Depth Perception
(1)	Constantly overshoots target, wide swings, slow to correct
(2)	
(3)	Some overshooting or missing of target, but quick to correct Accurately directs instruments in the correct plane to target
(4)	
(5)	**Bimanual Dexterity**
Score (1)	Uses only one hand, ignores non-dominant hand, poor coordination between hands
(2)	
(3)	Uses both hands, but does not optimize interaction between hands
(4)	Expertly uses both hands in a complementary manner to provide optimal exposure
(5)	**Efficiency**
Score (1)	Uncertain, inefficient efforts; many tentative movements; constantly changing focus or persisting without progress
(2)	Slow, but planned movements are reasonably organized
(3)	Confident, efficient and safe conduct, maintains focus on task until it is better performed by way of an alternative approach
(4)	
(5)	Tissue Handling
	Rough movements, tears tissue, injures adjacent structures, poor grasper control, grasper frequentlyslips
Score (1)	
	Handles tissue reasonably well, minor trauma to adjacent tissue(i.e. occasional unnecessary bleeding or slipping of the grasper)
(2)	
(3)	Handles tissues well, applies appropriate traction, negligible injury to adjacent structures
(4)	
(5)	

Table 2.2: Likert-scale technical skill perception questionnaire, from the five of the 6 domains in the GEARS assessment metric. Autonomy is left out, as no audio or video of the operating room is provided in our datasets.

Score	Depth Perception
(1)	Constantly overshoots target, wide swings, slow to correct
(2)	Some overshooting or missing of target, but quick to correct
(3)	
(4)	
(5)	Accurately directs instruments in the correct plane to target
Score	**Bimanual Dexterity**
(1)	Uses only one hand, ignores non-dominant hand, poor coordination
(2)	
(3)	Uses both hands, but does not optimize interaction between hands
(4)	
(5)	Expertly uses both hands in a complementary way to provide optimal exposure
	Efficiency
Score (1)	Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress
(2)	
(3)	Slow, but planned movements are reasonably organized
(4)	
(5)	Confident, efficient and safe conduct, maintains focus on task, fluid progression
	Force Sensitivity
Score (1)	Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage
(2)	Handles tissue reasonably well, minor trauma to adjacent tissue, rare suture breakage
(3)	
(4)	Applies appropriate tension, negligible injury to adjacent structures, no suture breakage
(5)	**Robotic Control**
Score (1)	Consistently does not optimize view, hand position, or repeated collisions even with guidance
(2)	View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant.
(3)	
(4)	Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant.
(5)	

1.3 Hand Movement Research

Extensive point-to-point hand movement research has been used to create models which correlate to skill in non-surgical settings [10, 11, 25]. Notably, the minimum jerk model has been used in the past with studies of stroke patients, to define the skill involved in a movement, for various weeks after recovery [3]. It was hypothesized that more skilled movements were smoother. It is known that when hand movements occur, there are submovements present which appear in the form of several individual movements occurring in sequence [10]. Potentially, smoother movements lead to more 'expert' behavior or performance.

3.1 Measures of Movement Smoothness

2.1.3.1.1 **Submovements** It has been observed that early movements in recover- ing stroke patients exhibit movement 'fragments', known as submovements, shown in Figure 2.4 [11]. These submovements have been used in several studies, to demonstrate different metrics for quantifying skill level in movement tasks. The main area of interest is a movement's smoothness, or the continuity and non-intermittency of a movement. This smoothness increases with learning and experience [26].



Figure 2.4: Typical movements(bold lines) and their extracted submovements(fine lines) from first and last days of a therapy session for a stroke patient [11].

Submovements may be used as a measure of how smooth a movement may be. More submovements that occur between two points should lead to lower overall smoothness and less ability to perform the task. Decomposing a movement into its' submovements is a nonlinear global optimization problem, to determine the correct parameters suitable to fit the submovements to smaller movements. An example of this is shown in Figure 2.4. A 'scattershot' algorithm has been introduced that finds the globally optimal submovements' composition, by calculating the probability of finding the globally best fit [11].

This scattershot algorithm uses support-bounded log normal(LGNB) curves, which can be produced with

$$B(t) = \frac{D(T_1 - T_0)}{\sigma\sqrt{\quad}} \exp\left(\frac{-1}{\quad} ln\ \frac{t - T_0}{\quad} - \mu\right)^2 \qquad (2.1)$$

Seattle Children's — Internal

$$2\pi(t - T_0)(T_1 - \qquad 2\sigma \qquad T_1 -$$

from $T_0 \leq t \leq T_1$, where D is the displacement resulting from the movement, T_0 is the movement start time, T_1 is the end time, μ controls skewness, or assymetry, and σ 13 determines kurtosis, or fatness, of the curve. This function was created by Palomondon, and describes the movements as originating from the sequential action of a set of velocity generators working in a cascaded fashion. Velocity profiles can be described by LGNB curves which are asymmetric and continuous [25].

2.1.3.1.2 Spectral Arc Length Measuring smoothness of a movement can be done with several different metrics, such as the number of peaks in the velocity profile, the minimum squared jerk profile, and the speed profile. The most recent and successful metric is called the spectral arc length (SAL) [27]. SAL uses the Fourier magnitude spectrum of the speed profile, and measures the arc length of the movement's normalized

Fourier magnitude spectrum. This is defined as:

$$\eta_{sal} = \int_0^{\omega_c} \left(\frac{d\hat{V}(\omega)}{d\omega} \right)^2 d\omega, \tag{2.2}$$

$$\hat{V}(\omega) = \frac{V(\omega)}{V(0)}, \tag{2.3}$$

where V(ω) is the Fourier magnitude spectrum of $v(t)$, and [0, ω_c] is the frequency band occupied by the given movement. ω_c is set to 40π rad/s, which covers normal human movement.

$\hat{V}(\omega)$ is the amplitude normalized magnitude spectrum [27].

When SAL was used to test the smoothness of recovering stroke patients, and compared with other popular smoothness metrics, SAL was the only metric used in which the smooth- ness improved as the patients progressed [3].

One downside of SAL is that it computes a higher smoothness for tasks that occur more quickly, instead of just analyzing the movement itself, shown in Figure 2.5. An improvement was recently made to SAL, called SPARC (SPectral ARC length), which is independent of temporal movement scaling, as shown in Figure 2.6 [28].

It should be noted that movement smoothness is highly task dependent, and should only be compared among similar tasks, with similar constraints. The kinematic data used to measure smoothness should also be a derivative of a position metric, like velocity or jerk. This is because the smoothness measures will highlight whether intermittency occurs in the movement and ignore features that are not changing with time.

Smoothness measures can have large differences depending on whether or not the measure is being recorded over a discrete movement, in which the movement goes from one finite point to another. If the movement is instead rhythmic, in which, for example,

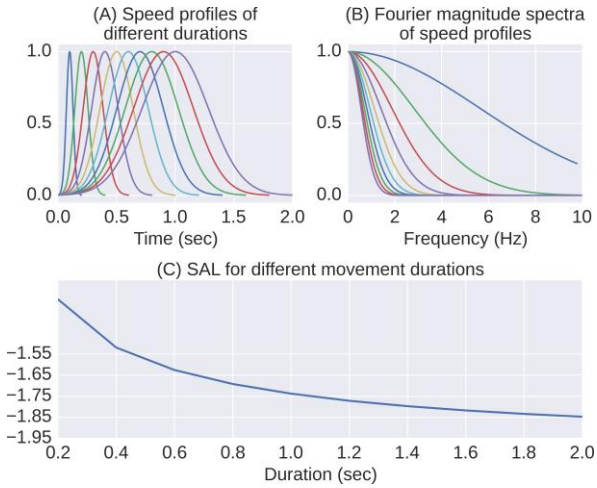


Figure 2.5: The effect of the SAL metric on varying lengths of movements [28]. Longer movements provide a lower SAL measurement.

a person's hand is moving back and forth in the same relative areas, repeatedly, a different method of measurement must be computed. If x(t) is the movement, then it can be represented as a concatenation of individual movements,

$$x(t) = \sum_{i=1}^{N} x^i(t); \; x^i(t) = x(t)\Pi_i(t);$$

$$\Pi_i(t) = \begin{cases} 1, & \text{if } t_i \le t < t_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

in which $x^i(t)$ is the ith movement, multiplied by the rectangular window $\Pi_i(t)$ of the ith movement. t_i is the start time of movement, t_{i+1} is the end of the movement. If a perfectly rhythmic movement is used, t_i can be chosen such that the concatenation
only captures certain portions of a movement.

The smoothness of each movement can then be found independently. In order to do

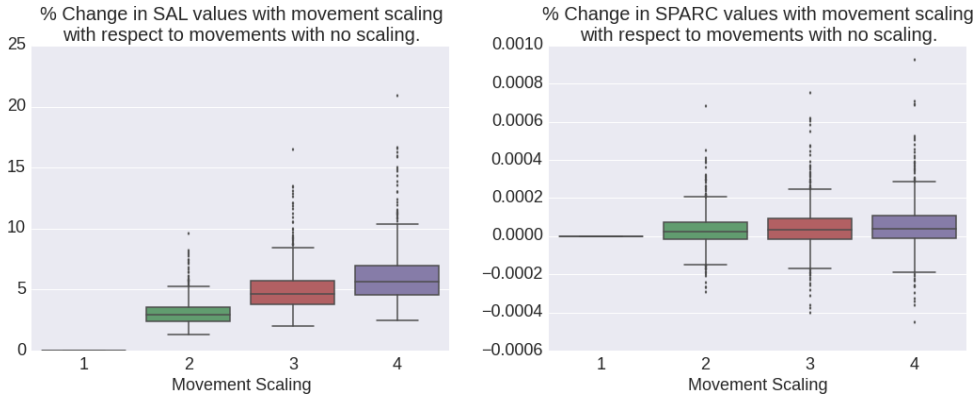


Figure 2.6: The effect of temporal movement scaling on SAL vs. SPARC [28]. SPARC is relatively impervious to movement scaling.

this, however, the amount of segments making up the total movement, N, should be known, and they should be at roughly equal lengths apart. These individual smoothness values can then be combined by use of a weighted mean function [28].

2.1.3.1.3 **Minimum Jerk** Research of velocity profile movements has been ana- lyzed in the past for predicting hand trajectories and finding relationships to perfor- mance or skill [12, 29–31]. [29] suggests that movements in which more skill is involved result in a large increase in spatial variability of the trajectory, as well as a slightly increased peak velocity.

It has been observed in [3] that if coordination is modeled mathematically, the qualitative and quantitative features of the movement may be elucidated. This can be done through the defining of a cost function, which here is the square of the movement's
jerk magnitude:

$$\frac{1}{0}\int^{t_f} \left(\frac{d^3x}{dt^3}\right)^2 + \left(\frac{d^3y}{dt^3}\right)^2 dt.$$

When Equation 2.5 is minimized, the smoothest possible movement of the hand is theoretically obtained. This optimally smooth movement also depicts a movement in which only one submovement exists. This study showed that unconstrained point- to-point movements result in a relatively straight path, and a bell-shaped tangential velocity profile, as shown in Figure 2.7.

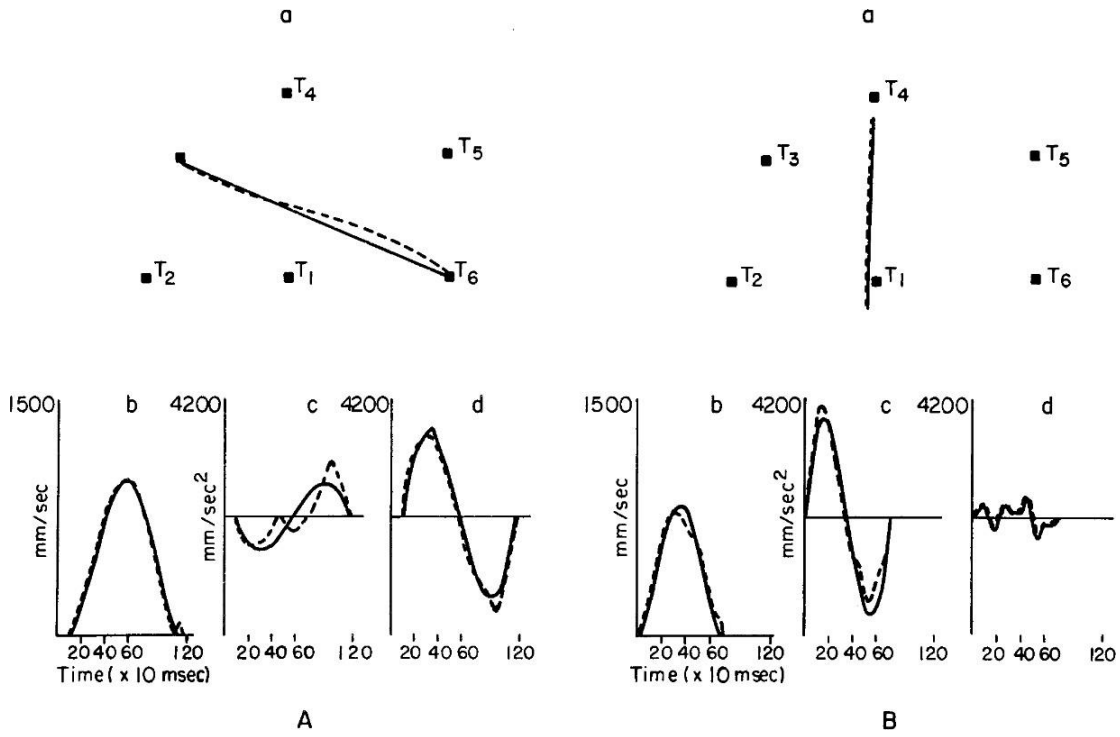


Figure 2.7: Predicted(solid lines) and measured(dashed lines) hand trajectories over- laid [3]. The minimum jerk model provides accurate predictions for the movements of recovering stroke patients.

Several models to predict and plot hand trajectories of unconstrained point-to-point movements exist, including the minimum jerk model, the support-bounded lognormal model, and the minimum snap model [25].

3.2 Objective

Given previous successes with neurophysiologically-derived models of hand movements, this thesis plans to test this with surgical data. This research will test the that the minimum jerk velocity prediction model will be able to accurately model the movements of expert surgeons, when compared to novice surgeons, using linear mixed effects modeling to account for variability of surgeons and task types.

1.4 Computational Skill Evaluation

Surgical skill evaluation is still a relatively niche field within research labs. Several different approaches exist with differing degrees of accuracy to compute surgical skill. The main approaches are using summary metrics calculated from surgical events and motion data, as well as statistical and machine learning models used from either kinematic data streams or video.

4.1 Neural Networks

Neural networks first were coined in 1943, when a McCullough theorized how neurons might work [32]. Since then, we have learned neurons do not work quite in this way but the term has stuck, due to the similarity to how the algorithms are computed. The first actual neural network was created in 1959 to eliminate echoes in phone lines [33]. Several groups have used neural networks(NNs) to help in the classification of images, videos, and other types of multidimensional data. There are several variations of neural networks, including:

- Convolutional NN (CNN)
- Recurrent NN (RNN)
- Gated Recurrent Unit (GRU)
- Long Short Term Memory Network (LSTM)

Neural networks use groups of ‘neurons’ or ‘perceptrons’, which make up one layer of a network, or a hidden layer, and feed data to remaining layers based on their activation function outputs. The input of a network is given to the neuron which computes one of several possible activation functions to assess the variablity of the inputs. Neural networks can be made of several layers, which are then called ‘deep neural networks’. A neuron in the hidden layer will receive several inputs from different input neurons, and the output of that neuron’s activation function is passed to the next layer. This continues until an output layer is reached, which will receive all the hidden layers’ in- formation and give an output based on that output layer’s activation function. Figure 2.8 illustrates this basic architecture.

Different types of NNs can be used for different types of problems. CNNs are excel- lent at recognizing patterns that repeat across data. This is why they are mostly used for face and object detection. Instead of every neuron being connected to the following layer, however, in a CNN the first ‘convolutional’ layer takes a small portion of, for example, an image, as input, one by one. The filter runs through the entire image, and the results are restructured into an array. Once all the images have been fed through the network’s pipeline, the algorithm is able to identify which objects are similar, and separates them into classes.

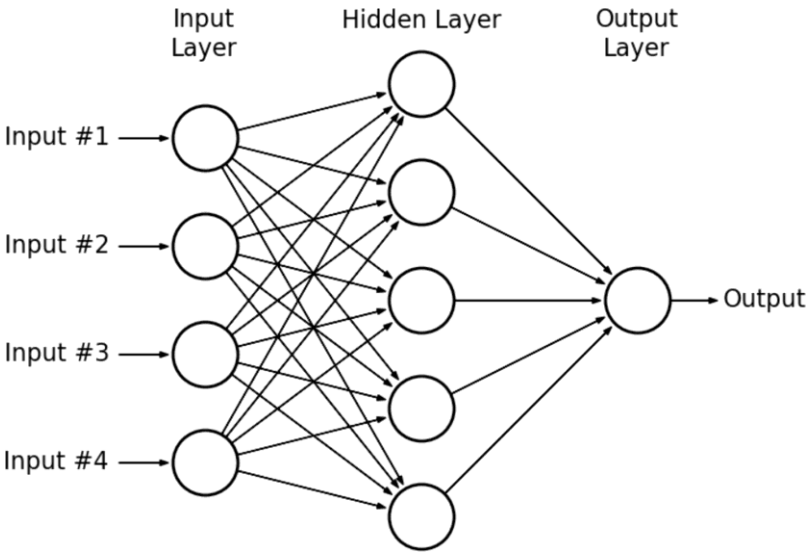


Figure 2.8: An example of the architecture of a neural network [34]. Inputs are sent through layers in which activation functions and array manipulations are used to arrive at a single output signal.

4.2 Recurrent Neural Networks

Recurrent Neural Networks (RNN) are different from CNNs in that they can take in arbitrarily-sized data and are useful for time-series information. This is possible because the output of the hidden layer in an RNN is fed back through the current layer with the previous layer's input. This allows RNNs to learn information from the past, as well as information from the present. RNNs are often used for language and vocabulary recognition, because they will 'remember' the context of a sentence, allowing prediction for certain words. RNNs are known to succumb to an issue known as the vanishing gradient problem. The vanishing gradient is a phenomenon susceptible to certain gradient-based optimization strategies in which a particularly small gradient is calculated during the optimization, and a domino-effect causes the algorithm to stop improvement due to no changes occurring in the gradient. Vanishing gradients can prevent a neural network from learning any useful information. This problem has been fixed, however, by long short term memory networks (LSTMs) [35].

4.3 Long Short-Term Memory

Long short-term memory networks are an even more recently created form of NN, which have longer dependencies than RNNs. In an LSTM, there are two inputs, a cell state and a hidden state. The cell state travels through a neuron to decide whether to dispose of the entering cell state or to store it. If the cell state is thrown away or forgotten, the hidden state makes an update to the cell state. The mechanism deciding whether to keep a cell state can vary between LSTMs, each having their pros and cons, but this is normally a mixture of tanh and sigmoid functions. An illustration of an LSTM cell is shown in Figure 2.9.

Seattle Children's — Internal

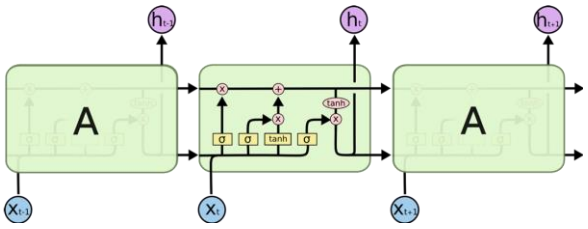


Figure 2.9: An example of the architecture of a long short-term memory network. A series of gates are used to create a sense of context for past mistakes and successes.

The most common way of setting up an LSTM is with a memory cell, input gate, output gate, and a forget gate. The terms are rather straightforward. The memory cell takes the input and stores it away, the input gate controls the likelihood of a new input flowing into the cell, the forget gate decides whether or not a value remains in the cell, and the output gate, controls whether the value in the cell is computed with the activation function. This activation function is usually the logistic function.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \qquad (2.6)$$

LSTMs have been used in speech recognition research with much success, and are part of the engine behind Apple's Siri speech recognition technology [36] [37], [38]. Speech models potentially provide promise for surgical skill evaluation due to the requirement of memory of past events. A speech model is required to keep a record of context to improve future predictions, just as a surgical skill classifier must record previous mistakes or successes to predict future skill levels.

4.4 Automated Performance Metrics

Automated performance metrics are summary statistics computed, most popularly with robotic surgery, as the daVinci surgical robot is able to track motion data through- out a surgical procedure. The daVinci robot, created by Intuitive Surgical (Sunnyvale, CA), also records events during a procedure, such as the number of times a surgeon moves their head in place of the console to conduct a movement, and the amount of times the hand or foot pedal is pressed on the console [39].

In [40] a combination of APM computation and task-evoked pupillary response (TEPR) was measured, to learn the affect of cognitive work load on the different skill levels. It was found that novice surgeons tend to move faster under high cognitive workloads, while experts move more slowly under the same cognitive experiences. They were additionally able to correctly classify the different skill levels using random forest classification with APM computation. It was found that the most reliable metric for cal- culating the skill of a surgeon was how often they used the camera clutching mechanism to move the camera while performing [41]. Other performance metrics exist relating to surgical skill which are discussed in more depth in Chapter 4.

4.5 Video Motion Data Analysis

Recent researchers have been able to use raw video data of tasks performed, and classify the skill of the performer [42–44]. These studies have used motion feature ex- tractors from image data to discriminate different types of motion representative of certain skill levels. This research has been fairly accurate, but have only been tested with small datasets, and for subjects wearing brightly colored gloves so the video can

differentiate certain movements from others. In a real-world setting the performer would not be wearing these brightly colored gloves, making it much more difficult for image detection to identify hands and classify motions.

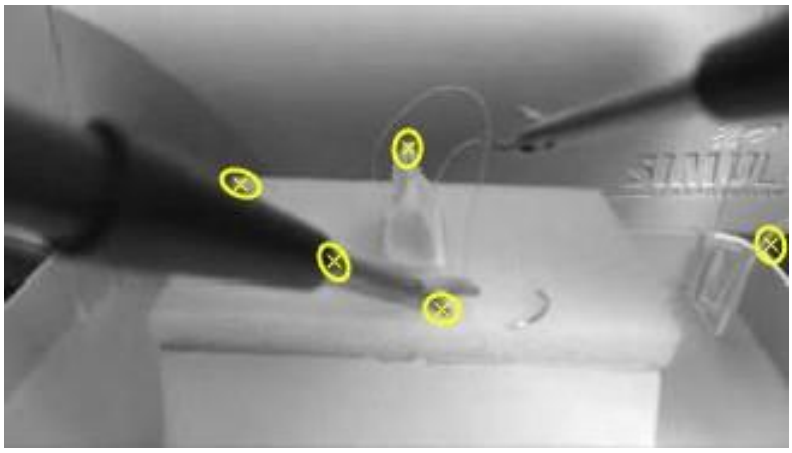
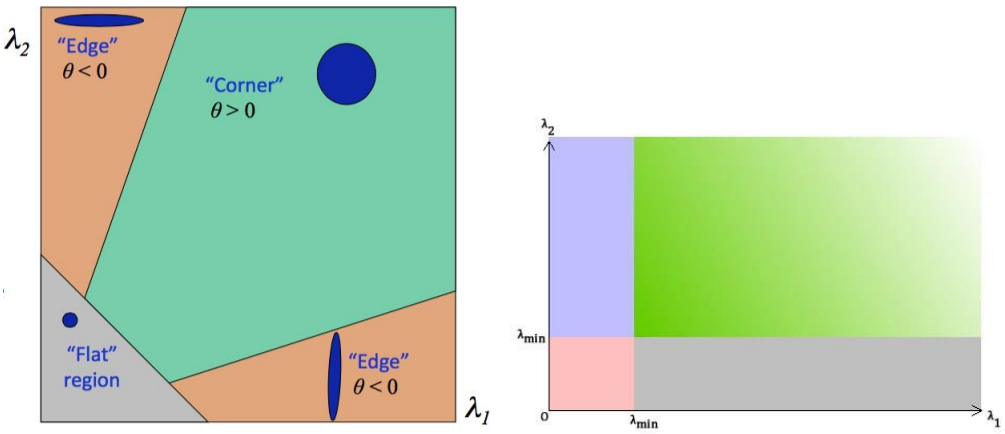


Figure 2.10: The detected SIFT points from a dry lab laparoscopic surgical task image (computed using MATLAB computer vision toolbox).

Several kinds of features may be extracted from video images. Some of the most popular methods are Harris corner detectors, Sped Up Robust Features (SURF), Scale Invariant Feature Transforms (SIFT) (shown in Figure 2.10, and, more recently, Spatio-Temporal Interest Points (STIP) [45]. These can all be used to help classify images in video, and are essentially small variations of basic corner detection [46], [47].

A more robust version of corner detection was created by Shi-Tomasi. This is a slight variation from Harris corner detection. Both methods calculate a score R, calculated for Harris detection by the equation



(a) The boundaries for harris corner detection to classify a point as a corner [49]

(b) The boundaries for Shi-Tomasi corner detection to classify as a corner [48].

Figure 2.11: Comparison of corner detection boundaries for Harris and Shi-Tomasi detection algorithms. Shi-Tomasi has a more rigid set of constraints, and will detect more corners because of these boundaries.

$$R = detM - k(traceM)^2$$

$$detM = \lambda_1\lambda_2 \; traceM = \lambda_1 + \lambda_2. \tag{2.7}$$

In Shi-Tomasi detection, the score is simply

$$R = min(\lambda_1, \lambda_2). \tag{2.8}$$

This small change has been shown to experimentally improve results. The change in boundary between classifying as a corner or an edge can be seen in Fig 2.11 [48].

These methods of extracting features consist of detecting corner points and subsequntly computing histograms of oriented gradients (HoG) in three directions (two

spatial and one temporal direction) [50]. These 3D HoG features will provide information for how the objects move through time as well as space. Several researchers have used variants of this technique for surgical skill evaluation [51–54]. One study went a step further and used these features to create frame kernel matrices.

A standard kernel matrix is an equation which gives a measure of similarity between two matrices. Using this with the features in image frames conveys the similarity be- tween image classes in subsequent frames. An example frame kernel matrix is shown in Figure 2.12. This approach used image features to divide the motions into 5 distinct motion classes using K-Means Clustering, and subsequently classify surgeons by which motion class most of their movements belonged [55]. This most publication reported an accuracy of 88%, making it the benchmark for other publications to compete with.

4.6 Lack of Quality Data

Most previous research in skill evaluation has suffered from a lack of robust data sets. The main studies completed with automated performance metrics were conducted with 26 cases. The most promising work which used video analysis of motion features used a data set of just 18 participants [56]. A summary table for some other popular studies with the size dataset are shown in Table 2.3. In order to create reliable models, studies with data of larger volumes must be used and made public to the research com- munity so that comparable results may be communicated.

Seattle Children's — Internal

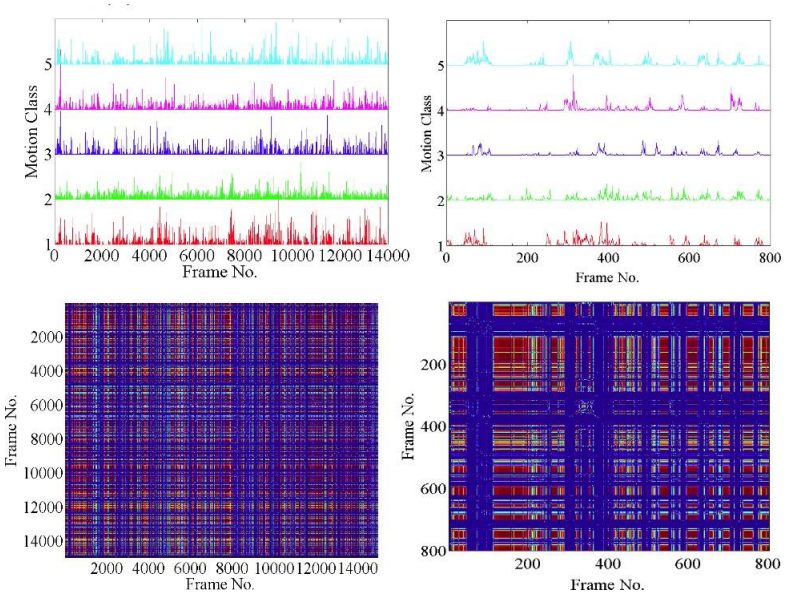


Figure 2.12: Top: Discriminated motion counts corresponding to each class of STIPs. Bottom: Frame kernel matrices example, displaying the relationship between various segments of a video [55].

Group	Technique	N
Chen(2018) [39]	APM	70
Nguyen(2019) [40]	APM	26
Zia(2017) [56]	Motion Analysis	18
Fard(2018) [57]	Motion Analysis	3
Sharma(2014)	Video Analysis	16
Zia(2017) [53]	Video Analysis	8
Funke(2018) [58]	Video Analysis	8
Wang(2019) [59]	Video Analysis	8

Table 2.3: Summary of popular skill evaluation studies, and lack of high volume data.

4.7 Objective

Given previous successes in fields unrelated to surgical skill evaluation in which time is a large part of the equation in correctly making predictions, it is believed an LSTM could be used in surgical settings. This thesis aims to create a machine learning model which uses an LSTM with tooltip kinematic data from dry lab laparoscopic training procedures to make predictions which will succeed in assessing the skill of novice and expert surgeons, surpassing previous research efforts summarized in this chapter.

1.5 Biological Motion Perception

As mentioned earlier in this chapter, task time seems to be the summative metric which most often correlates to surgical skill, as determined by experts or crowd scores of video evaluation [13]. As these are relatively subjective forms of measurement, they can be riddled with inaccuracies and biases that humans subconsciously obtain. It is possible that something which occurs in the brain causes humans to unconsciously reward quicker speeds when they are unaware of the speed up.

Seattle Children's — Internal

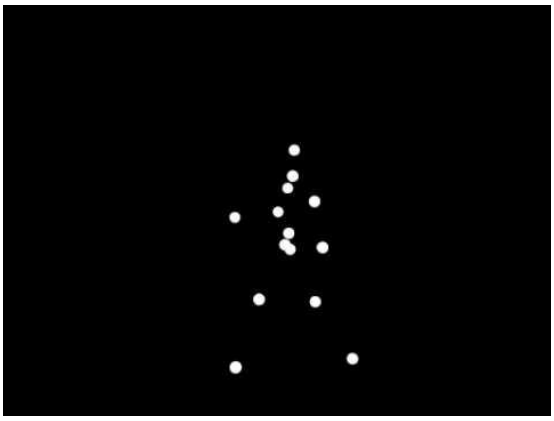


Figure 2.13: Example of a point light walker. Each of the dots represent a limb or joint of the human body.

5.1 Visual Stimuli Research

Psychophysical experiments have tackled the question of sensory stimulus using different methods of manipulation. The first of these is a study in which animations of point light (PL) animations were used, illustrations of moving dots representing human joints (Fig. 2.13) [60]. This concluded that observers can classify gender from one of these animations with 63% accuracy. Additionally, manipulations of unnatural arm swinging, different walking speeds, as well as removal of different joints significantly reduced the ability to recognize gender, to nearly 50%. Interestingly, faster walkers did appear more female to observers, although to a non significant degree [61]. This work was continued in a similar study of PL animations, finding that two revolutions of gait are needed to accurately determine gender [62], and another finding that only one-tenth of a second is needed to recognize human motion [63]. This work shows that motion perception is somehow linked to dynamics. Our brains appear to have an ability to effortlessly decode this information.

Seattle Children's — Internal

5.2 Response to Abnormal Dynamics

It has been shown when these various animations are manipulated to move at abnormal or unnaturally slow speeds, instead of perceiving slower movement, observers perceived the animation to be rotating [64]. In contrast, when animations are played at speeds outside the realm of human possibility using spatiotemporal filters, observers also become inefficient at recognizing the gender [2]. Likewise, [65] found that humans are able to accurately assess which speeds appear natural. It is possible that biological motion perception has a speed dependence in which the speed occurring must appear to be realistic for the motion perception unit to properly process it.

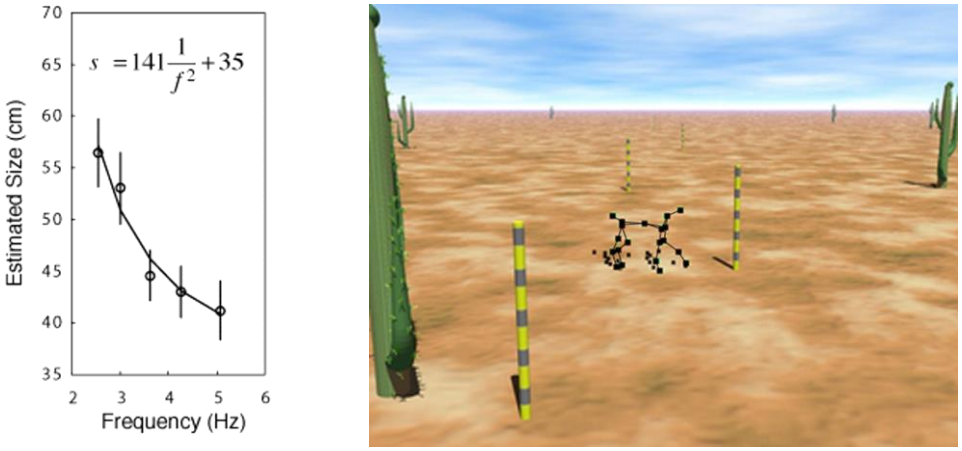


Figure 2.14: Illustration of a PLW representing an animal (right) and the change in perceived size as the gait increases (left). This suggests a link between speed and biological motion perception [66].

5.3 Objective

Driven from previous research in biological motion perception, it is hypothesized that humans will perceive the skill of a surgeon differently when video playback speed is increased. There may exist a biological motion processing unit in the brain which is able to detect movement which seems ‘unrealistic’ when moving at speeds too fast to

attribute to normal human speeds. Once this threshold is met, I hypothesize the brain will begin to control for the increase in speed and this perception of increased skill will cease. This infers a bell-shaped unimodal function of perceived skill as video playback speed is increased.

1.6 Datasets to be used in this work

6.1 Basic Laparoscopic Urologic Study

The Basic Laparoscopic Urologic Skills(BLUS) dataset is made up of 535 videos recording four different laparoscopic tasks: peg transfer, pattern cutting, suturing, and clip applying. This data was gathered from 8 urologic training centers in the United States from 76 unique subjects who ranged from 1st to 5th year residents. There are 49 R1 subjects, 11 R2 subjects, 6 R3 subjects, 6 R4 subjects, and 4 R5 subjects. The videos have: 133 peg transfer, 127 pattern cutting, 135 suturing, and 140 clip applying tasks. These data include both videos of the task and tool motion data. Included are also expert and crowd-sourced ratings for each trial [19]. Frames from three of the four tasks are shown in Figure 2.15, with the range of scores for each of the four tasks in the dataset shown in Figure. 2.16.

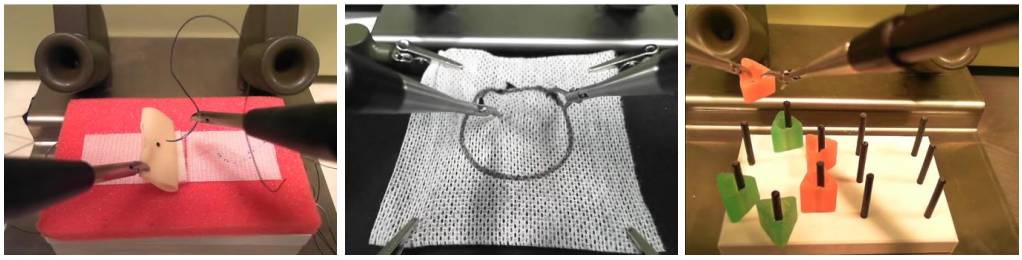


Figure 2.15: Frames from a (left to right) suturing, cutting, and peg transfer task in the BLUS dataset.

6.2 Robotic Surgery Readiness Study

The Robotic Surgery Readiness Study data set is made of 343 videos of live robotic surgery cases which all hold accompanying kinematic tooltip data logging the tool mo- tion during each video. This data set arose through research aiming to test whether

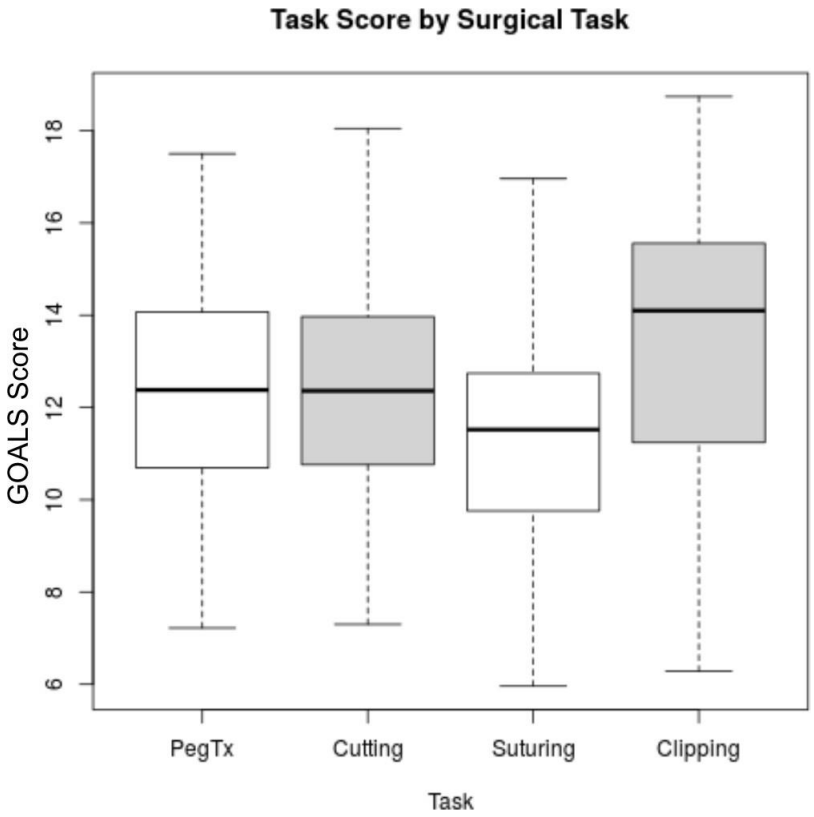


Figure 2.16: Range of crowd-assessed surgical skill evaluation mean scores for each of the BLUS tasks. On average clipping tasks receive higher scores than, for example, suturing tasks.

pre-operative warm-up using a validated simulation curriculum improved robotic surgical performance among practicing surgeons with lapses in practice. These data are explained in much more detail in Chapter 6 as they were created during the past three years. An example figure from the RSR dataset is shown in Figure 2.17.

1.7 Linear Mixed Effects Modeling

When data is grouped into several subgroups in which a relationship may exist, a useful mechanism to use is a linear mixed effects (LME) model. The assumptions made

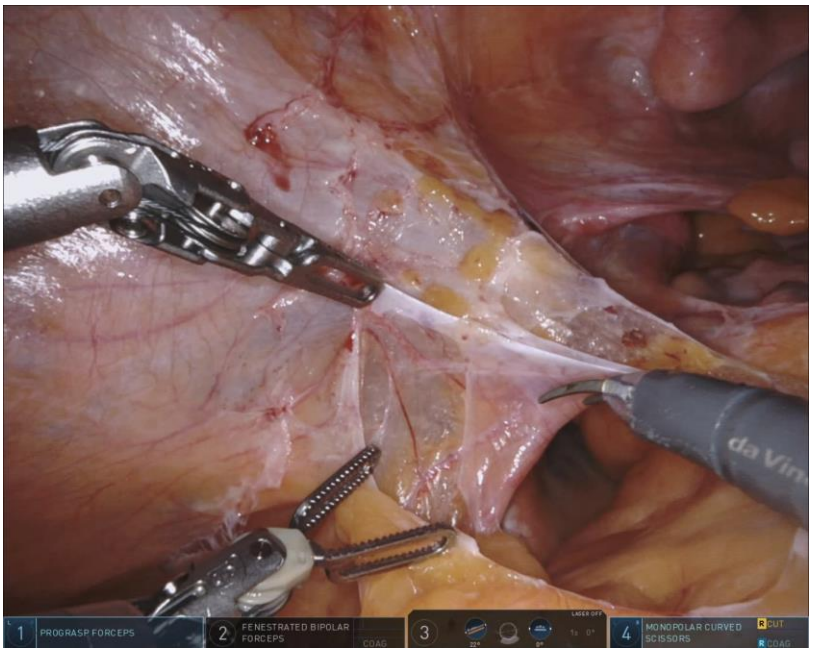


Figure 2.17: Frame from a video in the RSR data set. Typical robotic surgeries recorded were prostatectomy, hysterectomy, and partial nephrectomy, with all procedures being recorded in Washington state.

before using an LME model is that the data is not independent and thus we can use a combination of fixed and random effects to create an advanced linear regression which will explain the hierarchical structure of the data. One example of this sort of data may be several doctors seeing a variety of patients who either have or have not taken a drug [67].

Fixed effects are parameters that do not vary throughout the dataset, and are as- sumed to be measured without error. This could be something like the amount of a drug taken for specific groups of patients. These variables are easily repeatable in mul- tiple studies. In contrast, a random effect is something that changes throughout the experiment, such as a separate doctor seeing each patient. In this way we can account for different doctors having different baseline levels at which they may make certain

decisions, as well as each doctor having a different range of symptoms they would like to see before making certain decisions. Modeling random effects in this example gives each doctor a specific slope in the regression, and each fixed effect is given a specific intercept in the regression. By modeling our data this way we can account for the variability of fixed effects, things measured without error, but also have variables which account for the factors which changed throughout the experiment. This general setup in mathematical form would be

$$y = X\beta + Zu + E \quad (2.9)$$

in which the outcome y is a vector of size Nx1, X is a Nxp matrix of the p predictor variables, β is a px1 column vector of the fixed-effects coefficients, Z is the Nxq design matrix for the q random effects, u is a qx1 vector of the random effects, and E is a Nx1 column vector of residuals, or errors not explainable by the model [68].

When testing a linear mixed effects model, in order to learn whether a certain pa- rameter was significant in the regression, the standard method to use is an ANalysis Of VAriance (ANOVA) test. The typical setup is to use the null model, the model which does not include the parameter of interest, and test with a model which includes an effect for the desired parameter. The information criteria from an ANOVA test can then give a sense for whether the parameter is significant in predicting an outcome by comparing the information criteria. If the information criteria for the new model is lower than the original, as well as having a reasonable p-value, this confirms the parameter is significant in predicting the outcome variable.

Chapter 3

Neurophysiological Models

This chapter provides the methods used to test whether velocity prediction models are correlated to surgical motion segments.

3.1 Testing the Relationship of Minimum Jerk Model to Surgical Skill

1.1 Methods

The already obtained velocity trajectories from the BLUS dataset were segmented using an algorithm which tracked when the tooltip velocities fell to near zero. Addi- tional forms of segmentation have been used, including presence of a tool grasp in each motion, and motions corresponding to the force of a grasp passing a certain threshold, but the most statistically significant form of segmentation is for segments separated at near-zero velocity segments.

The speed segments from each of the performances in each of the four BLUS tasks

Seattle Children's — Internal

were fit to a minimum jerk velocity profile, and the mean squared error of the differ- ence between the points on the fitted curve and the raw data were recorded for each movement and recorded as belonging to the surgeon who made that specific movement. In an effort to make the data more manageable, these error values were used for the remaining calculations.

Initially a fixed effects model was used, which accounted for a fixed effect in each β coefficient, but did not account for the error of each specific surgeon. The fixed effects model was of the form

$$TaskScores \sim \beta_1 + \beta_2 + \beta_3 + \beta_4 + MSE * Subjects, \quad (3.1)$$

in which each of the β values are coefficients used to fit the minimum jerk model to the specific hand movement, and MSE is the mean squared error of the fitted line to the model. Additionally, a linear mixed effects model was used which included a random effect for each of the subjects and tasks. An ANOVA test was then used to compare the null model to a model which included the error term. The model being tested was

$$TaskScores \sim (1|Subjects) + (1|Task) + MSE. \quad (3.2)$$

This equation allows for variability within Tasks as well as within specific performers of the tasks. This allows the model to know each task will have a different range of scores, as well as knowing each surgeon will have a baseline level of performance.

1.2 Results

Using a surface plot which allows for plotting with various values of the alpha transparency coefficient, the velocity of each segment of each performance of Peg Transfer

tasks was overlaid onto one figure, with the time component normalized. As can be seen in Figure 3.1, a vague shape emerges from the overlaid segments showing the general trend of grasping movements. Figure 3.2 shows a visualization of the segmented grasps for the first observation of the peg transfer tasks, with normalized time and velocity segments.

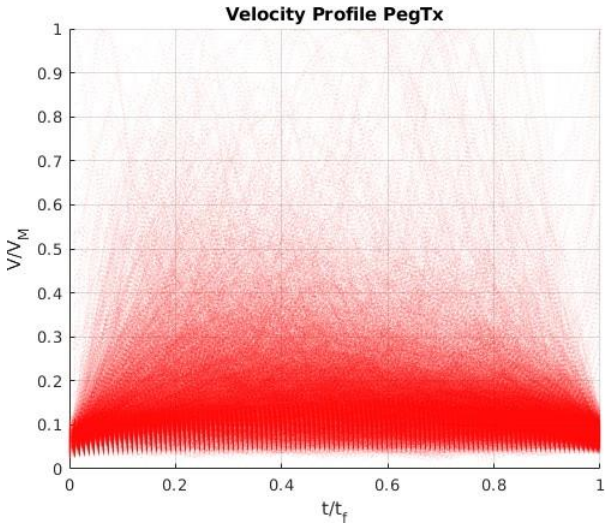


Figure 3.1: All Peg Transfer task segments overlaid to show the general shape of peg transfer task grasp segments. One single line represents one single movement by a surgeon. Movements were defined as starting and ending when the hand reached nonzero speed. Time is normalized by the duration of each segment and velocity is scaled by the peak velocity magnitude of that segment.

Figure 3.3 shows an example of the difference between the minimum jerk model's fit to grasps and the raw grasp data. Figure 3.4 show the results of the fixed effects model in Equation 3.1, with the actual vs. predicted scores plotted. These show a correlation of 0.521.

This result for the model in Equation 3.2 is shown in 3.5 and has a correlation of 0.826 to the outcome scores. To test the viability of this model, an ANOVA test was used

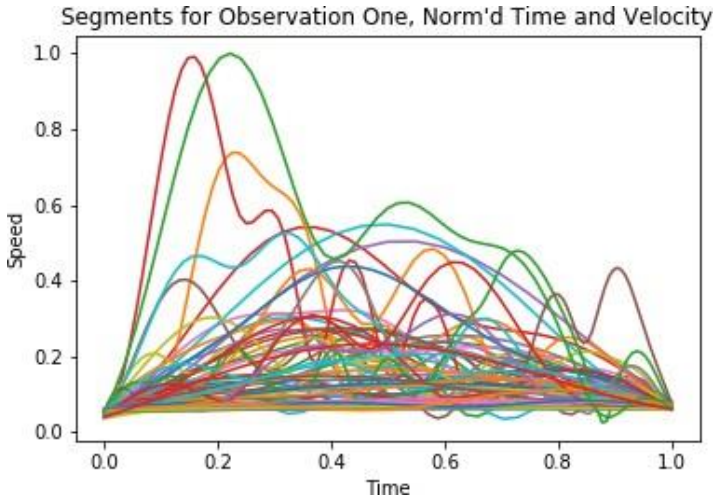


Figure 3.2: Example of segmented grasps, from the first performance of the peg transfer tasks only. Showing the data in this manner allows seeing the variability that exists throughout the different movements in a surgical task performance.

to compare the linear mixed effects model shown above, to the null model, excluding MSE, as that is the variable of interest. The results are shown in Table 3.1. The AIC, or Akaike Information Criterion, is a well-known measure of the relative goodness of fit of a model, when being compared to a similar model. This AIC can be calculated by

$$AIC = 2k - 2ln(\hat{L}) \tag{3.3}$$

in which k is the number of parameters, and $\hat{L}$ is the maximum likelihood. When comparing two or more models, the model which has a lower AIC is preferred. It is standard practice to use AIC to compare a model to the null model, which excludes the variable of interest.

As can be seen, the AIC is lower for the model with an MSE term included, and the χ^2 value is 951.45 with a p-value of ≤ 0.0001. From the summary of the LME model,

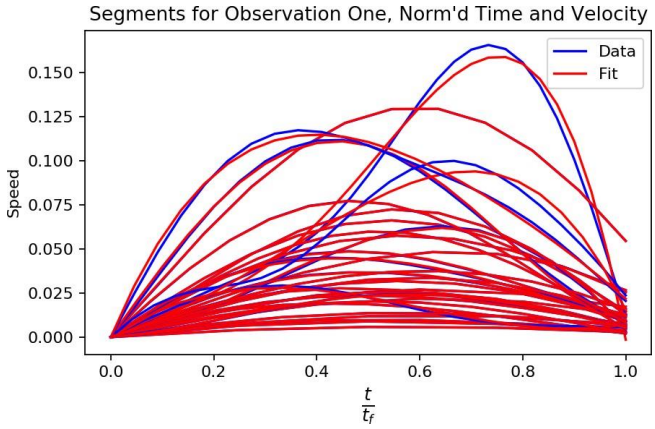


Figure 3.3: Unimodal movements from peg transfer tasks fitted to minimum jerk move- ments described in Chapter 2.1.3.1.3.

the MSE fixed effect term raises the predicted score of a surgeon by 0.4324 ± 0.2803. This suggests a higher error actually leads to a higher score, or that higher scores mean a worse fit to the minimum jerk model.

The LME model, plotted separately into the four tasks performed, can be seen in Figure 3.6.

Model	AIC	BIC	logLik	devian ce	Chisq	Chi Df	P- Value
Null	74390. 18	74421. 87	- 37191.0 9	74382. 18	NA	NA	NA
MSE Model	73448. 73	73520. 03	- 36715.3 6	73430. 73	951.44 89	5	$\leq$ 0.0001

Table 3.1: ANOVA test results for the mixed effects model on the minimum jerk viability as a predictor variable for task score value.

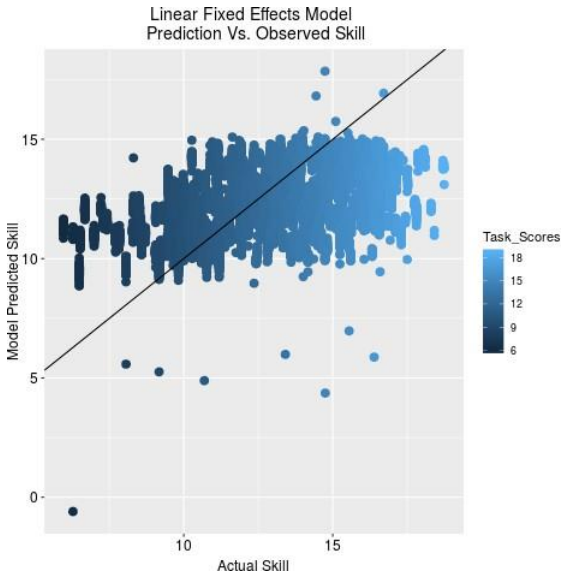


Figure 3.4: Fixed effects model of observed vs. predicted skill, for all BLUS tasks combined, using minimum jerk model, subject and task identity, in which the fixed effects accounts for a constant intercept for all of these factors. The fixed effects model behaves quite poorly.

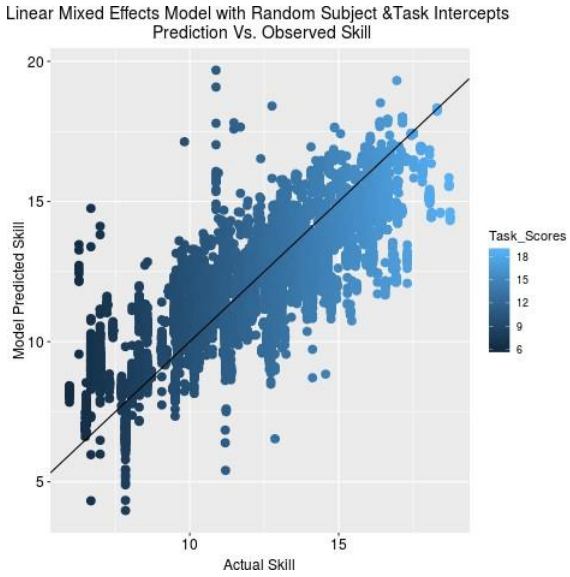


Figure 3.5: Linear Mixed effects model's predicted vs. observed skill using minimum jerk model fit error, task, and subject identity, in which this model accounts for varying intercept values for each task, and subject. This model is able to predict outcomes much better than the fixed effects model, with a correlation between actual and predicted skill of 0.826.

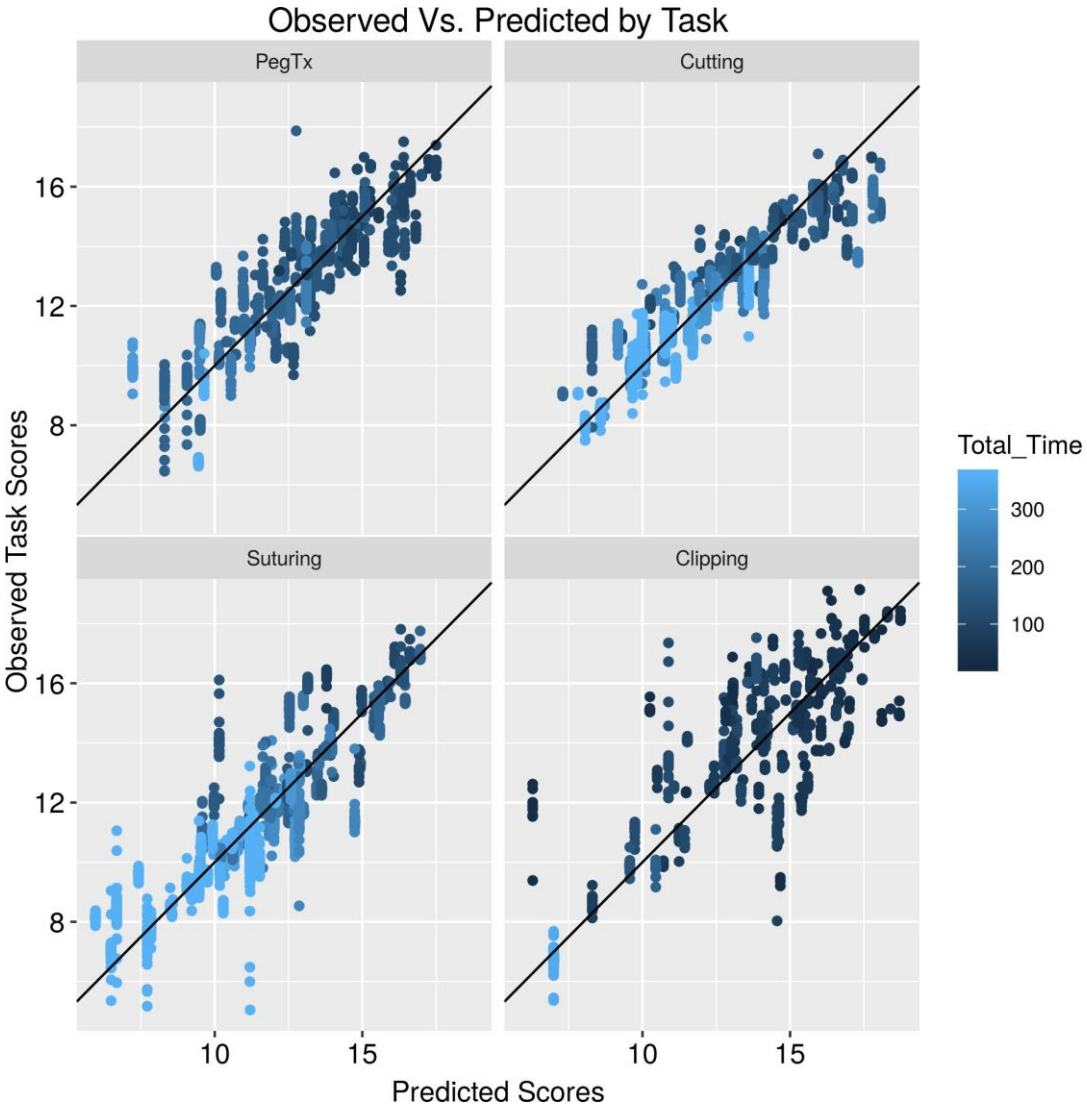


Figure 3.6: LME predictions grouped by task, with colorbar illustrating total time for task completion. There is a strong correlation between observed task scores and predicted scores, with slower performances usually receiving poorer of both.

Chapter 4

Computational Skill Evaluation

This chapter will discuss efforts in creating a machine learning model for use with multidimensional kinematic data extracted from the BLUS dataset. This model was used to classify the skill of surgeons performing various laparoscopic training procedures. This is a reproduction of a journal paper submitted to the International Journal of Computer Assisted Radiology and Surgery in March 2020 [69].

Seattle Children's — Internal

4.1 Bidirectional Long Short-Term Memory for Surgical Skill Classification of Temporally Segmented

Tasks [69]

1.1 Abstract

Purpose: The majority of historical surgical skill research typically analyzes holistic summary task-level metrics to create a skill classification for a performance. Recent advances in machine learning allow time series classification at the sub-task level, al- lowing predictions on segments of tasks, which could improve task level technical skill assessment.

Methods: A bidirectional long short-term memory (LSTM) network was used with 8 second windows of multidimensional time-series data from the Basic Laparoscopic Urologic Skills (BLUS) dataset. The network was trained on Experts and Novices from four common surgical tasks. Stratified cross validation with regularization was used to avoid overfitting. The misclassified cases were re-submitted for surgical technical skill assessment to crowds using Amazon Mechanical Turk to re-evaluate, to analyze the level of agreement with previous scores.

Results: Performance was best for the suturing task, with 96.88% accuracy at predicting whether a performance was an expert or novice, with 1 misclassification, when comparing to previously obtained crowd evaluations. When compared with expert surgeon ratings, the LSTM predictions resulted in a Spearman coefficient of 0.89 for suturing tasks. When crowds re-evaluated misclassified performances it was found that for all 5 misclassified cases from peg transfer and suturing tasks, the crowds agreed more with our LSTM model than with the previously obtained crowd scores.

Conclusion: This technique shows promise both for classifying performances of surgical tasks and evaluating sub-task level segments of performances. This method could be used for near real-time feedback to surgeons. Further work will be done to expand this methodology to a larger cohort of surgical performances including robot- assisted surgery.

1.2 Introduction

Computationally assessing the skill of a surgeon in an objective manner using tool motion has proven a complex problem with many challenges. Previous research has re- lied mostly on summary performance metrics from kinematic data [41], [13, 27]. Unfor- tunately, these metrics typically failed to completely discriminate novices from experts, that is to never misclassify “obvious” novices vs. “obvious” experts – the so-called Min- imally Acceptable Classifier (MAC) Criterion [70]. Recent advances in machine learning techniques have expanded the possibility of using time series datasets to evaluate surg- eries and surgical tasks [71].

The de facto gold standard for determining the level of skill in a surgical performance is video-based evaluation by an expert surgeon [72]. Skills assessment is normally eval- uated with the use of one of several possible established assessment schemes that utilize Likert-scale scoring of subdomains relevant to skill, such as bimanual dexterity, or tissue handling. An example of one of these assessment methods, the Global Operative As- sessment of Laparoscopic Skills (GOALS) evaluation scheme is shown in Table 2.1 [24]. This is both a time-consuming and laborious process for experienced surgeons whose valuable time could rather be spent performing life-saving surgeries. In recent years it was found that crowd based evaluation of surgical

Seattle Children's — Internal

performance is able to achieve the same standard of accuracy in predicting technical skill in surgery, with a 100% pass/fail rate, as compared to expert surgeons [22]. This may still be a subjective measure of technical skill and prone to bias [73], thus an objective technique to computationally evaluate the technical skill of a surgeon would be beneficial.

Seattle Children's — Internal

Long short-term memory networks (LSTMs) are an adaptation of recurrent neural networks which are capable of analyzing past events in a time series to learn how they might affect a present time index [35]. This is possible through a series of gates in the architecture of the network which hypothesize, learn, and forget predictions by deduc- ing what information to ignore and which to emphasize in the training process. These networks have been further improved through the implementation of bidirectional net- works [74]. These behave by training a network in the forward direction, while also training a network in the reverse direction, thereby connecting two hidden layers in different directions to create one output from twice as much information, and allowing the LSTM to use this increase in information to achieve better results.

The objective of this investigation was to evaluate the importance of temporal seg- mentation of surgical tasks for quantifying the skill of a surgeon. The hypothesis of this work is that experts don't behave in an expert-like manner throughout the entirety of a task, and likewise for novices, but instead that the main factor in deciding upon a surgeon's overall technical skill is the number of expert-like to novice-like segments in footage of a surgical task, and that bidirectional LSTMs have the ability to learn this information from kinematic tool motion data.

1.3 Methods

3.1 Dataset

This study used the Basic Laparascopic Urologic Study (BLUS) dataset, described in detail in [19], with a summary re-iterated here for convenience. This dataset arose

from a gap in the field, in which no educational surgical certification process existed for urologic surgery, as opposed to how the Fundamentals of Laparosocopic Surgery (FLS) exists for general surgical procedures [75–77]. The BLUS training curriculum aimed to address urology appropriate skills improvement by recording video performances in an initial validation project of over 450 videos [78].

This dataset contains 454 videos of surgical performances consisting of four surgical tasks (110 peg transfer, 110 pattern cutting, 115 suturing, 119 clip applying), which are performed by medical students, urology residents, fellows and faculty surgeons from eight academic urology training centers in the United States [79]. Each trial of a surgeon performing one of the four tasks was recorded at 30 fps with a fixed camera-position of the laparoscopic tools interacting with the training field. Each trial additionally has kinematic data, sampled at 30 Hz, logging the tooltip positions, grasping force, and the jaw angles during the performance, as well as demographic information for each performer being obtained. A GOALS score was obtained for each video via crowd eval- uation, and an expert evaluation was obtained for a subset of videos.

Previous research regarded suturing the most clinically relevant of the four tasks, as these performances require the mastering of needle and suture handling which are more similar to what is encompassed in real surgery vs. transferring synthetic blocks or gauze cutting. However, all four tasks were used in this study in an attempt to provide a classification scheme which can successfully separate experts from novices. Here a novice was defined solely as an “obvious novice”, or someone who should never be allowed to operate and experts solely as “obvious experts”, or surgeons who should never be disqualified from operating [70]. An obvious expert was chosen such that the performer was in the top 20% of previously obtained GOALS scores for that particular BLUS task. The obvious novices were chosen in the same fashion such that they were in the bottom quintile of these domains. This method aimed to provide two well dis- criminated clusters of skill levels that should demand no misclassifications.

3.2 Skill Classification from Temporal Segmentation

3.3 Data Partitioning

In the preliminary analysis, a sliding window data partitioning technique was used in which windows of varying sizes with varying overlapping lengths were tested to find the optimal parameters. The network was trained by using the novices and experts from the top and bottom quintile in each of the four tasks such that the classifier could more easily find separating features to define each class. These values are shown in Table 4.2, where the minimum possible GOALS score is 4, and the maximum is 20. These labels are then converted into binary variables based on whether they are considered an expert or novice. This results in 16 expert videos and 16 novices videos for peg transfer, suturing and cutting tasks, and 16 expert videos and 17 novice videos for the clip applying task.

3.4 Window Parameter Selection

Due to the relatively small dataset resulting from sub-selecting only obvious novices vs. obvious experts, a stratified cross validation scheme was computed on the experts and novices, in which two performances were left out of training. Each of these groups of two performances consisted of an expert and a novice, in an effort to keep the training response values at equal proportions and prevent overfitting of the dataset. In addition, to avoid oversampling of novice performances, as novices consisted of about 70% of the total data, being that novices usually take longer to perform, the expert-labeled performances were sampled with a step size half as large as the novice step size, to simulate more training data. This sliding window and the specified sampling control technique is illustrated in Fig. 4.1.

3.5 LSTM Parameters and Architecture

The bidirectional LSTM network consisted of a one layer 32 unit bidirectional cell, followed by a 50% dropout layer, and a sigmoid activation function. L2 regularization was also used to further prevent overfitting. Each network was trained to 100 epochs using a binary crossentropy loss function, and the Adam optimizer. Each segment of a test performance was fed to the network and evaluated. After all segments were eval- uated, the LSTM outputs a probability of how likely a segment is to be expert-like vs. novice-like. The task as a whole was considered an expert level performance if the mean of the predictions resulted in a prediction of greater than 0.50, and vice versa.

After the initial results were obtained from using experts and novices as training and test data only, additional testing was done using intermediates as test data on these models. During this trial, the same training technique was used as discussed previously, this time alternatively using intermediate level performers as the test data. This training was repeated ten times per validation set, with different random number seeds being assigned at the intialization of each trial. For each intermediate perfor- mance prediction, the ten different predictions were averaged for each segment of the performance. An example of this is shown in Fig. 4.2.

Using these intermediate performance prediction values, it was now possible to get a

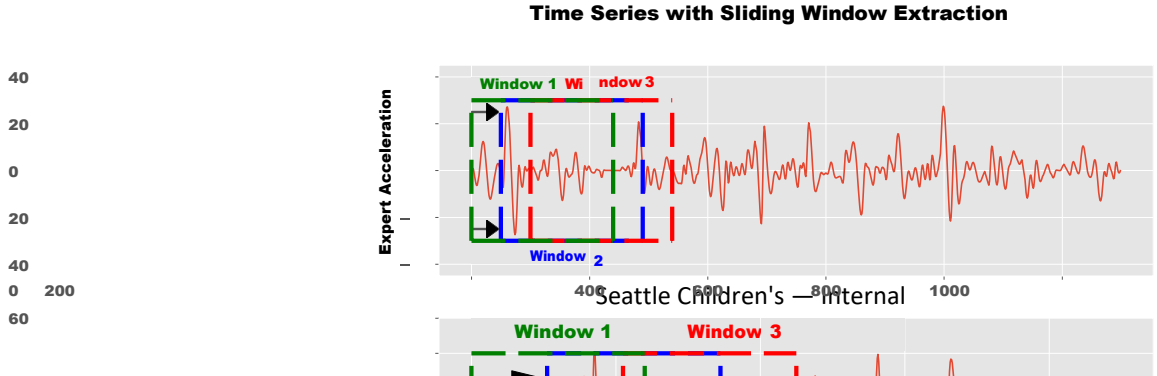


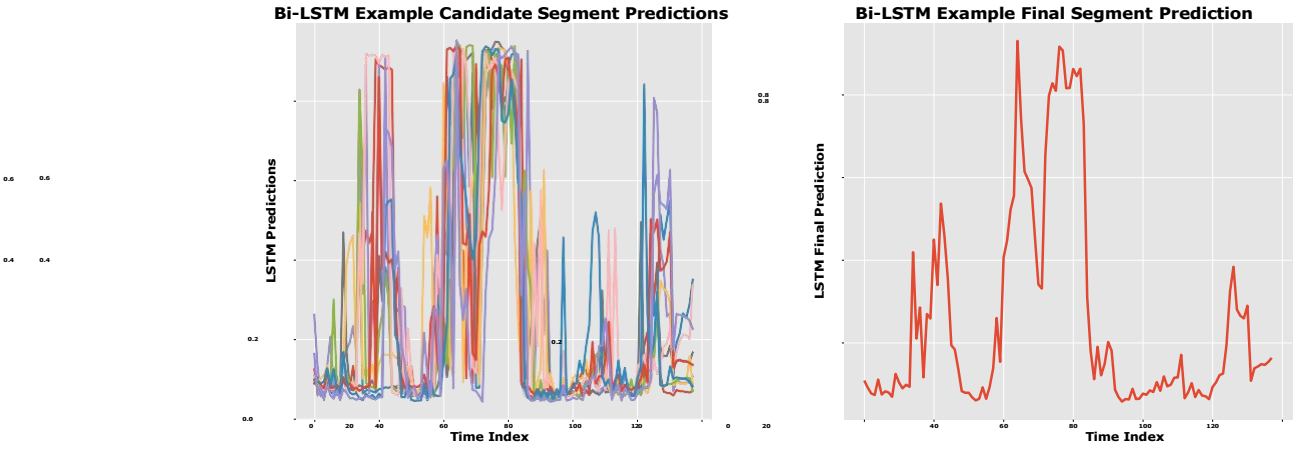


Figure 4.1: Example left hand tool acceleration time series with the sliding time window extraction method applied to both a novice series and an expert series. A novice perfor- mance is indexed with a window overlap of 105 time indices, and an expert performance is indexed with a window overlap of 50 time indices.

rough approximation of the correlation between the previously obtained GOALS scores and the average of the performance's LSTM predictions. These values were obtained for the experts and novices, as well as the intermediate performers.

3.6 Crowd Reassessment

As suturing and peg transfer tasks are agreed to be the two most separable and clinically relevant tasks from the BLUS dataset, these two were chosen to have their misclassified videos reassessed by crowd workers. In addition 5 obvious experts and 5 obvious novices from each task were randomly selected for reassessment. Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each



(a) 10 Test Predictions for a suturing perfor-mance.

(b) Final averaged prediction with outliers removed.

Figure 4.2: Process of computing final suturing performance prediction. All ten candidate predictions of a performance are averaged at each index to compute the final prediction.

non-expert crowd worker was paid an average of $0.50 to watch and evaluate the short video of the surgical task. The goal was to compensate the evaluators at a rate of approximately $10/hour. A user interface was created which asked crowds to rate the skill level of the performance using the GOALS assessment method, as obtained previously. These videos were given to crowds one at a time, using 40 crowd workers per video. The mean of the ratings for each video was then taken and compared to the previously obtained ratings for the misclassified videos, to find the level of agreement.

1.4 Results

4.1 Skill Classification from Temporal Segmentation

After performing a grid search with different hyperparameter values, the network was found to perform optimally with a window sample size of 240 time indices, approx- imately equal to 8 seconds, with a 3.5 second overlap at each window for novices and

Drop out	Batch Size	L2 Regularization
0.1	120	0.1
0.2	**240**	**0.2**
0.4	360	0.3

Table 4.1: Main hyperparameters tested during model evaluation in Section 3.1, with most optimal results in bold.

approximately a 1 second overlap for expert performances, which results in the number of samples from experts and novices to be roughly equal. Most of the values tested are shown in Table 4.1.

Table 4.2 illustrates the accuracy of these networks in classifying overall skill in surgical task settings, achieving over 96.88% accuracy for the suturing task, which is usually seen as the most clinically relevant task. Table 4.2 also reports the expert and novice specific accuracies, showing that all of the novices for suturing were correctly classified. Of the 32 suturing videos labeled as experts or novices, only one was misla- beled. By getting the average of all predicted segments in a specific performance, the algorithm results in a number between 0 and 1, codifying the LSTM's prediction of skill for the performance. These predictions can then be used to arrive at a correlation coefficient signifying the degree of correctness in the neural network at evaluating in- termediate performers in addition to experts and novices, when comparing to crowds, illustrated for each BLUS task in Fig. 4.3. Correlation values for each task are in Table 4.3. Suturing had the highest Pearson correlation coefficient with a value of 0.86.

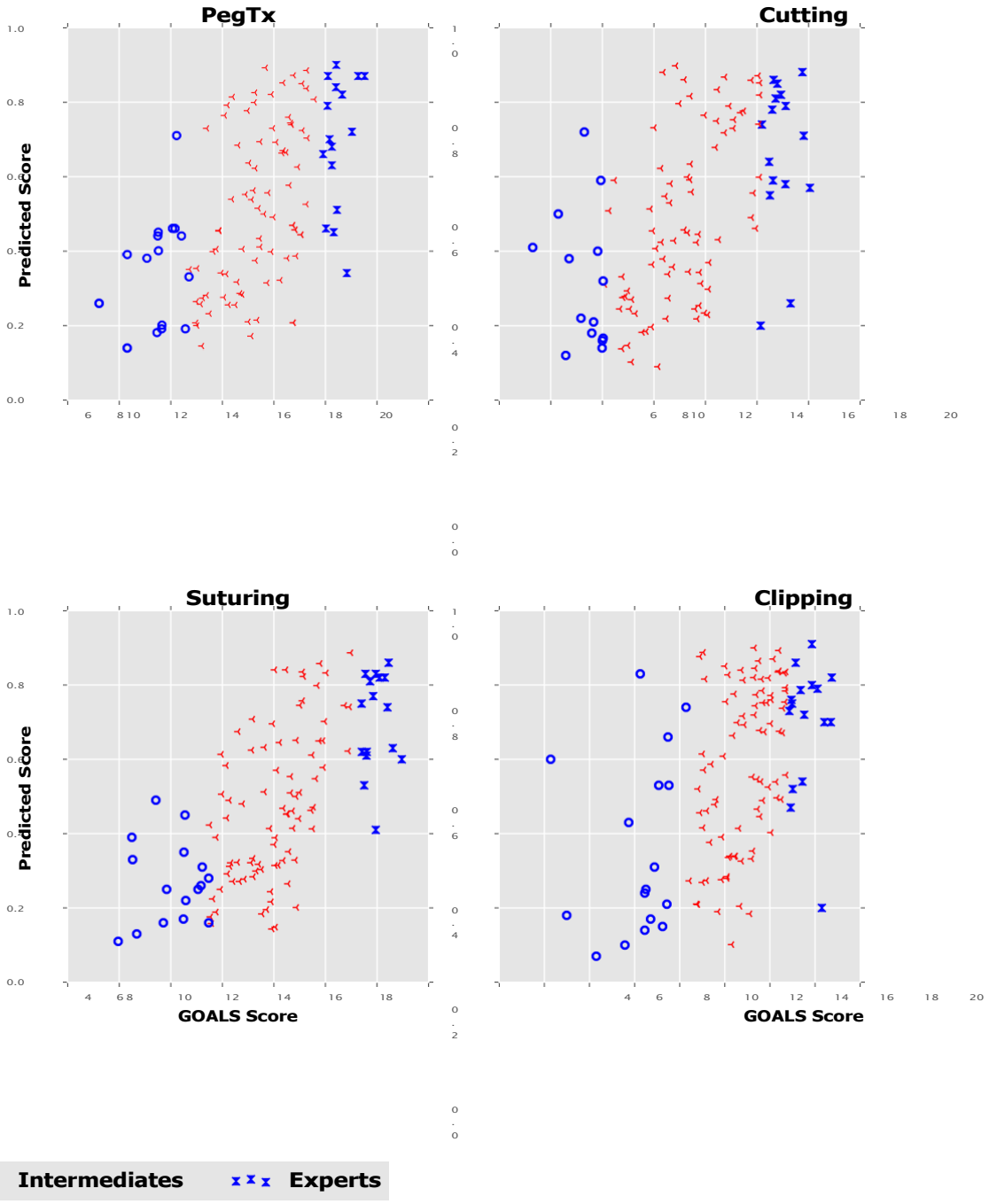


Figure 4.3: The LSTM's prediction for all performances from each of the tasks in the BLUS dataset. The LSTM was only trained on experts and novices (summary score labels with cross validation), from Section 3.1.

The suturing task was the most accurate of the four tasks in the BLUS dataset. Twelve of the suturing performances were previously rated by expert faculty surgeons in addition to being evaluated by crowds. Fig. 4.4 shows the faculty scores plotted against the skill predictions obtained from the bidirectional LSTM. As can be seen,

Task	Accuracy	Expert Threshold	Novice Threshold
Suturing	**96.88 %**	15.40+	9.47-
Peg Transfer	87.50%	15.89+	10.69-
Cutting	87.50%	16.14+	10.03-
Clip Applying	73.33%	16.86+	12.28-

(a) Accuracy for each BLUS task's experts and novices, from Section 3.1.

Task	Novice-Specific Acc.	Expert-Specific Acc.
Suturing	**100%**	**93.75%**
Peg Transfer	**93.75%**	81.25%
Cutting	87.50%	87.50%
Clip Applying	68.75%	87.50%

(b)Novice and expert specific accuracy for each BLUS task, also known as sensitivity and specificity.

Table 4.2: Accuracy results computed by the bidirectional LSTM for each BLUS task.

Task	Spearman	Pearson
Suturing	0.76	**0.86**
Peg Transfer	0.72	0.76
Cutting	0.61	0.69
Clip Applying	0.60	0.63

Table 4.3: Correlation coefficients for the relationship between GOALS scores and pre- dicted scores from the LSTM for every performance from each of the four BLUS tasks, including intermediate level performers, from Section 3.1.

there is a strong positive correlation, which has a Spearman coefficient of 0.89 and 91.67% accuracy.

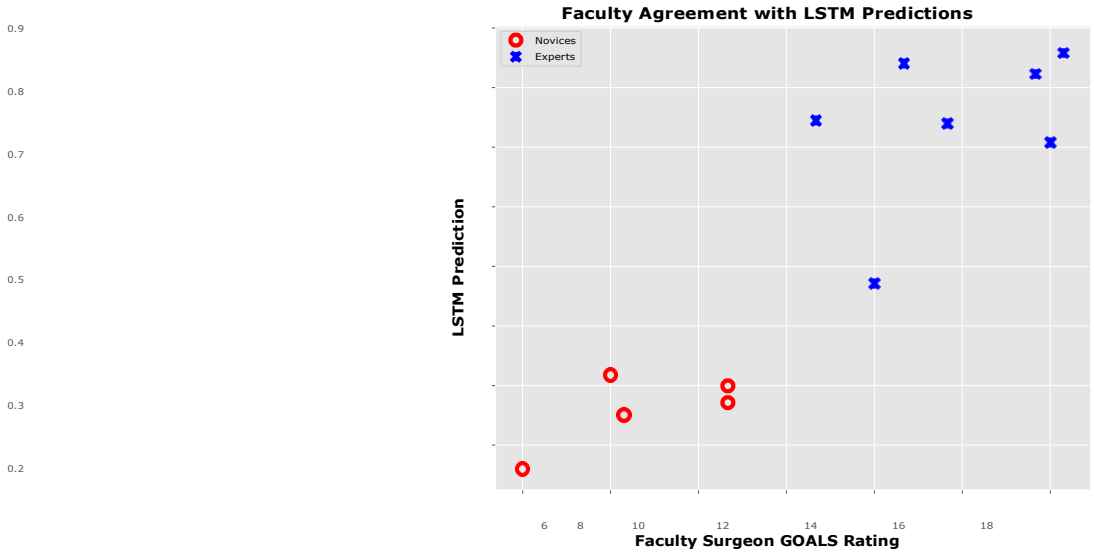


Figure 4.4: The LSTM predictions trained on crowd scores, for performances of suturing tasks as rated by faculty surgeons, which has a Spearman correlation of 0.89, from Section 3.1.

Once the model is trained with this kinematic data, it doesn't only allow classification of experts and novices, but in addition, allows the output to show the model's perceived skill of the surgeon through time for each of the subsequent segments, on a range of 0 to 1. One may view a video of a peg transfer task, and view how the LSTM model perceives the score of the surgeon as time progresses. It can be seen in one of these examples, that as the user makes mistakes, the perceived score decreases, and once they fix their mistakes and speed up, the LSTM is able to perceive their score increasing. A single frame example of this is shown in Figure 4.5.

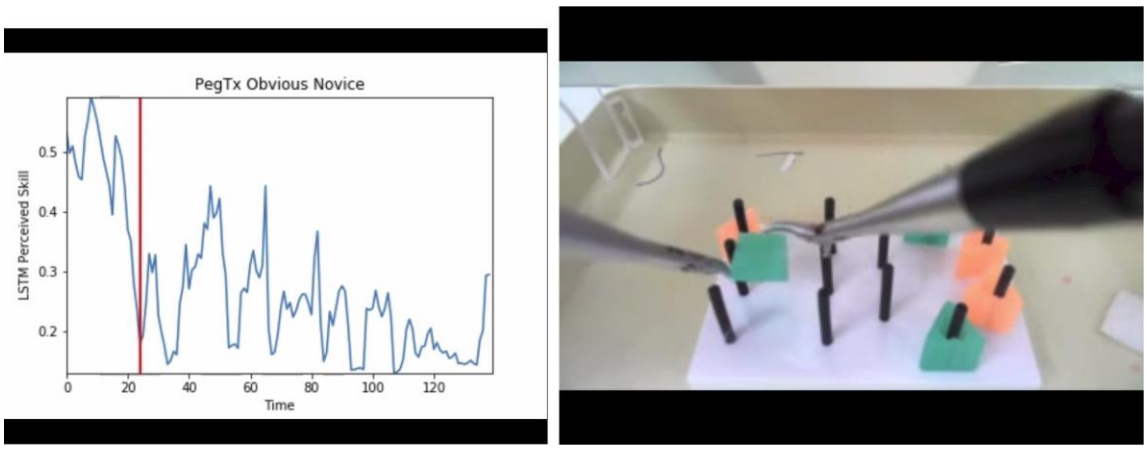


Figure 4.5: The perceived score of the LSTM on the left, with a video on the right of the same task.

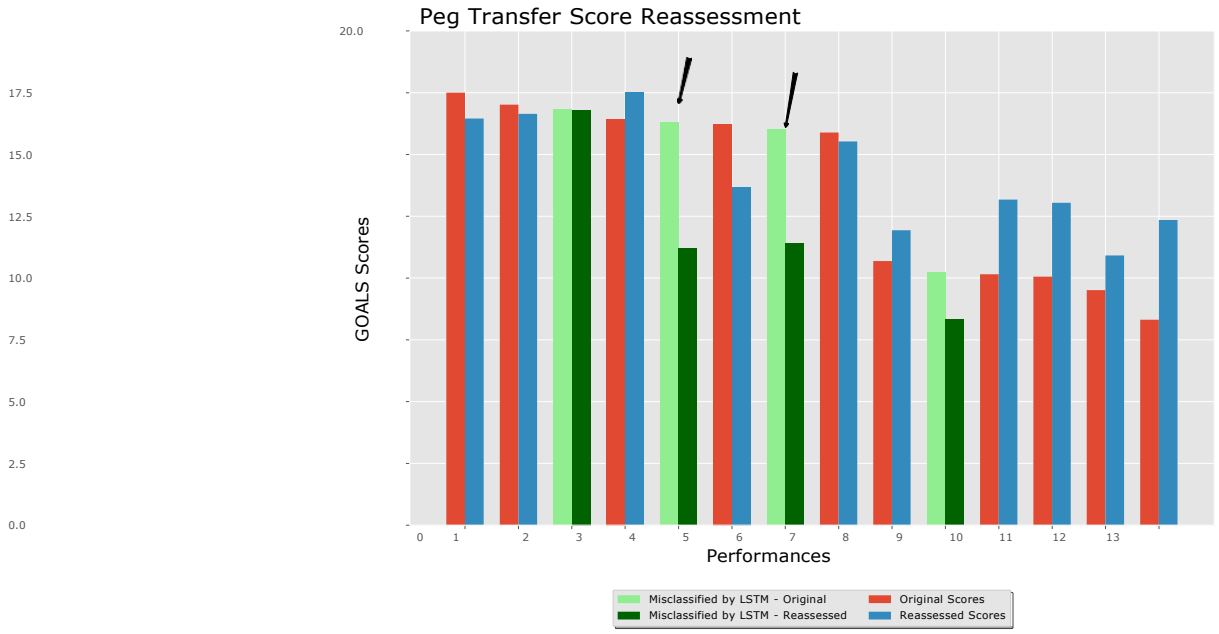
4.1.4.2 Crowd Reassessment

The suturing and peg transfer tasks were the two tasks with both high levels of classification accuracy as well as having a stronger correlation between scores and pre- diction levels, with a total of five misclassifications among the two tasks. The performances which were misclassified by the algorithm, which was trained on previously obtained crowd scores, were re-assessed by crowds. In addition to those performances, randomly selected obvious novices and 5 randomly selected obvious experts were additionally chosen for reassessment.

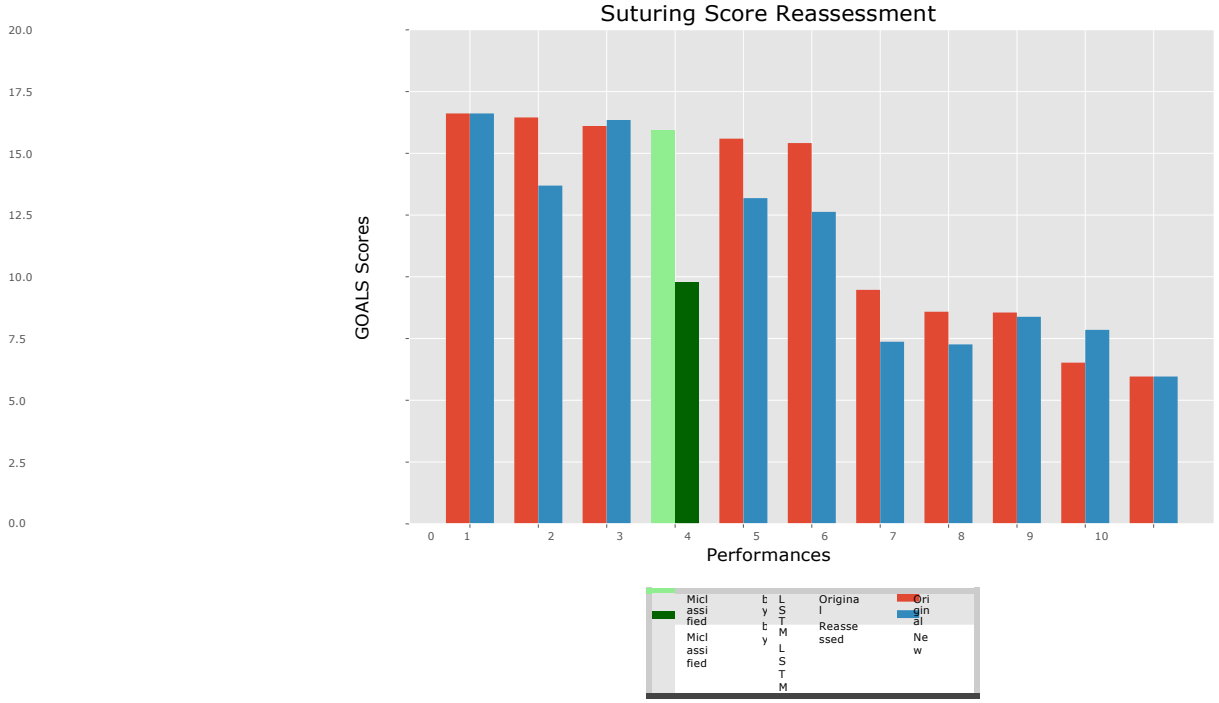
Surprisingly, the range in the reassessed scores was quite lower than in the original assessments. This could have been caused from a variety of reasons, such as the newer user interface used as well as having had each video individually evaluated separately compared to the previous method in which videos were evaluated in batches. However, the reassessed ratings do still have a general agreement in ranking of performances as the original scores. Interestingly, if the new evaluations are normalized to be in the range of the old scores (5.96-16.61 for suturing and 8.31-17.5 for peg transfer), 3 of the 5 performances which were misclassified by the LSTM were reassessed to no longer be classified as having the skill level initially given to that performance, as shown in Figures

and 4.7. Figure 4.7 illustrates the propensity of the reassessed crowd scores to agree with previously obtained scores or to agree with the LSTM, based on the skill level predicted by the algorithm, and whether there was a misclassification. These figures suggest that the LSTM's classifications for those 3 performances could have been more accurate than the original ratings given to those performers.

Seattle Children's — Internal



(a) GOALS scores of misclassified peg transfer tasks before and after reassess- ment, as well as 5 obvious experts and 5 obvious novices, chosen randomly. Two of the 4 misclassified performances were reassessed to be more in line with what the LSTM predicted, indicated by the arrows.



(b) GOALS scores of misclassified suturing tasks before and after reassess- ment, as well as 5 obvious experts and 5 obvious novices, chosen randomly. The only misclassified performance was reassessed to be more in line with what the LSTM predicted.

Figure 4.6: Reassessment of misclassified suturing and peg transfer performances suggest crowds agree more with LSTM than previous crowd ratings, from Section 3.2.

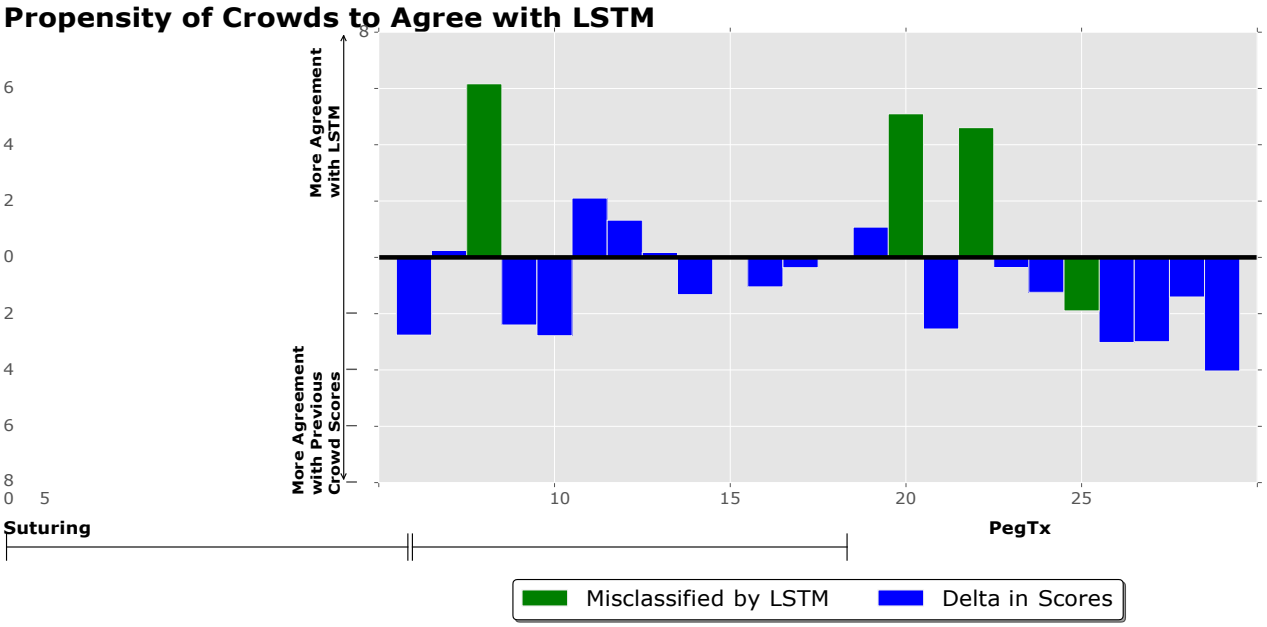


Figure 4.7: A bar plot of the magnitude of attenuation of the crowds to agree more with the LSTM on score reassessment, than with the original scores obtained.

1.5 Conclusion

The results from the bidirectional LSTM (Section 3.1), provide evidence in support of our hypothesis, that novice and expert surgeons do not exhibit expert performance metrics continuously throughout a task (and similarly for obvious novices). This suggests that temporal segmentation of kinematic tool motion analysis could provide more informative feedback of the skill of a surgeon, as compared to static summary performance metrics. The results, which also show good correlations between predicted score and actual score for several intermediate level performers, provide some additional evidence against overfitting, since these tasks were held out during the training phase.

Other popular methods of evaluating surgical skill such as computer vision algorithms which train on the frames of video, or popular automated performance metrics [41] could possibly be combined with this technique to create an even better classification model. This would further alleviate concerns about tool motion alone lacking important context that is still present in the video. Future iterations could enable giving correct predicted ratings of a performance, even during the surgical task performance, leading to near real-time skill feedback.

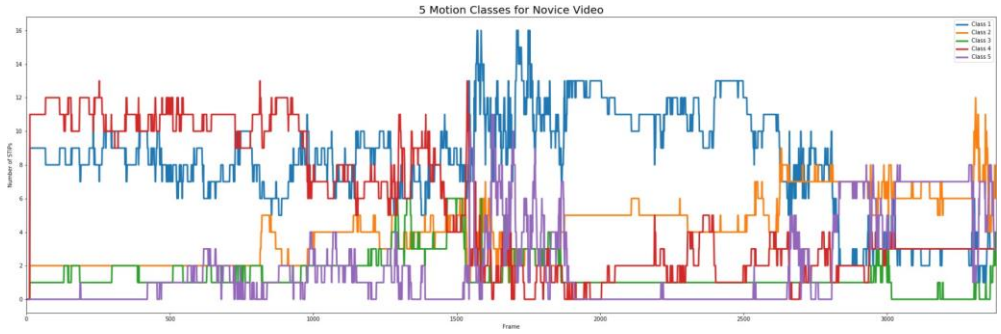
This study included some limitations. First and foremost, these tasks are simulated procedures, and the proposed algorithms and techniques may perform differently on real surgeries. We intend to test these hypotheses in the future on robot-assisted surgical data obtained from practicing surgeons. The proposed method also only analyzes tool motion data, which may not contain sufficient data required for complete skill classification [80]. The need for the reassessed scores to be normalized in order to have a comparison of crowd scores could be due to the techniques used in obtaining crowd evaluations which were different than the original evaluation methods. The authors acknowledge the results from the crowd reassessment could be in part due to differences in the user interface of the evaluation webpages used for the two different occurrences of testing.

42 Classification Using Video Data

Each of the videos from the Peg Transfer task were split into frames, and Shi-Tomasi corner detection points were computed for each. From here, the histogram of oriented gradients was computed by finding the derivative in both the x-direction, y-direction, and the temporal-direction, to find how these points change over time and space to get a sense of the optical flow. These points were then used in a similar way to how Zia used them to evaluate surgical skill [42]. Two expert videos of these points were concatenated to make one matrix, and a K-Means Clustering algorithm was used to divide the points into ‘motion classes’ which well symbolize the motion classes for expert motion in a peg transfer task.

Different numbers of motion classes were used to find the optimal number of classes to discriminate skill levels in video. Figures 4.8 and 4.9 show the motion classes in a sample video of a novice and expert.

Figure 4.8: The frames of a novice video with the amount of points belonging to each motion class being shown.



These matrices can additionally be split into time windows which can be run through a Gaussian kernel function to visualize the similarity between frames in a certain time window. For example, in Fig. 4.10 we can

Seattle Children's — Internal

see a frame kernel matrix for a portion of a novice video.

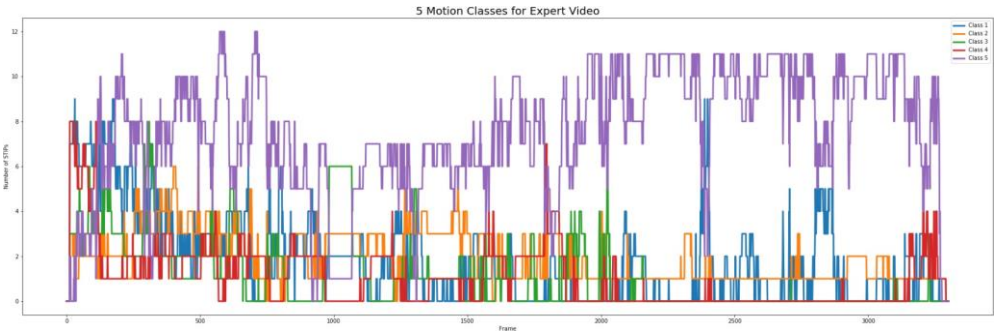


Figure 4.9: The frames of an expert video with the amount of points belonging to each motion class being shown.

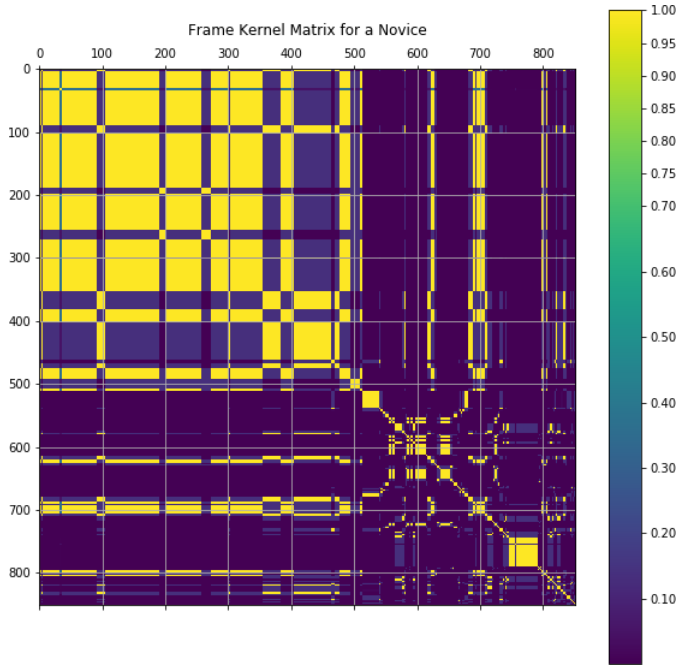


Figure 4.10: Frame kernel matrix for a portion of a time series, to visualize similarity between frames.

After obtaining the motion classes, the data was used as the input for the previously created bidirectional LSTM model, as well as a standard neural network model. Both of these models gave very poor results, with accuracy in classification reaching less than a coin toss. This proves to us that as it stands the amount of granularity in kinematic tooltip data is several degrees more complex than what is available in the video frames. The machine learning model is able to hone in on the complex details of the performance, the surgeon's movements, which most correlate to surgical skill. Using these algorithms with the frame-by-frame video data prevents the algorithm from being able to properly learn features when the algorithm is also being fed images which include much background information that isn't necessary. As it stands, it seems the best method of correctly classifying surgical technical skill from machine learning methods is by using tooltip kinematic data from which a model can most easily parse out the data. This method is able to currently outperform state-of-the-art machine learning methods with an accuracy of over 95% for tasks which closely replicate actual surgical performances.

Seattle Children's — Internal

Chapter 5

Speed Perception

This chapter discusses research regarding human speed perception and discusses experiments examining the existence of a bias in evaluating surgeons with different ratings when viewing videos of training exercises or surgeries at various speeds. The following is a reproduction of two separate publications: a dual journal and conference publication accepted to the International Journal of Computer Assisted Radiology and Surgery as well as the International Conference on Information Processing in Computer-Assisted Interventions at Munich, Germany in June 2020 [73], and a journal publication to the International Journal of Computer Assisted Radiology and Surgery, submitted in April 2020 [81]. Both of these publications contain their own abstract, introduction, methods, results, and conclusions, allowing each to stand alone.

5 The Effect of Video Playback Speed on Surgeon Technical Skill Perception [73]

1.1 Abstract

Purpose: Finding effective methods of discriminating surgeon technical skill has proven a complex problem to solve computationally. Previous research has shown that obtaining non-expert crowd evaluations of surgical performances is as accurate as the gold standard, expert surgeon review. The aim of this research is: (1) to learn whether crowd-sourced evaluators give higher ratings of technical skill to video of performances with increased playback speed, (2) its effect in discriminating skill levels, and (3) whether this increase is related to the evaluator consciously being aware that the video is manually manipulated.

Methods: A set of ten peg transfer videos (5 novices, 5 experts), were used to evaluate the perceived technical skill of the performers at each video playback speed used (0.4x-3.6x). Objective metrics used for measuring technical skill were also computed by manipulating the corresponding kinematic data of each performance. Two videos of an expert and novice performing dry lab laparoscopic trials of peg transfer tasks were used to obtain evaluations at each playback speed (0.2x-3.0x) of perception of whether a video is played at real-time playback speed or not.

Results: We found that while both novices and experts had increased perceived technical skill as the video playback was increased, the amount of increase was significantly greater for experts. Each increase in the playback speed by 0.4x was associated with, on average, a 0.72-point increase in the GOALS score (95% CI: 0.60-0.84 point increase; $p < 0.001$) for expert videos and only a 0.24-point increase in the GOALS score (95%

CI: 0.13-0.36 point increase; $p < 0.001$) for novice videos.

Conclusion: Due to the differential increase in perceived technical skill due to in- creased playback speed for experts, the difference between novice and expert skill levels of surgical performances may be more easily discerned by manually increasing the video playback speed.

1.2 Introduction

Medical errors make up for a third of all deaths in the United States, one of the largest contributors of which are surgical errors [4]. Technical surgical skill is directly related to patient outcomes [72], but it remains a difficult computational task to correctly classify surgeons into skill levels with a compelling level of accuracy, i.e. never misclassifying an 'obvious novice' as an 'obvious expert' and vice versa - the MAC Criterion [70]. The de facto gold standard for evaluating technical skill is video evaluation by an expert surgeon using Likert-scale assessment metrics, in which evaluators submit ratings on an anchored scale of 1-5. Using crowds of non-expert evaluators are a surprisingly accurate way to inexpensively and rapidly obtain skill level ratings for videos of surgical performances, with a pass/fail rate capable of matching 100% of ratings by expert surgeons [22]. The fact remains, however, that humans can be biased in their thinking, and subjective metrics of rating performances can lead to results we would no expect from computational models of evaluation.

One of the most popular laparoscopic surgical skill assessment metrics is the Global Operative Assessment of Laparoscopic Skills (GOALS), which is the most common objective assessment tool for laparoscopic performance studies [24], [82]. The subdomains in this metric include: bimanual dexterity, tissue handling, efficiency, depth perception, and autonomy. Task time is not a direct metric used to evaluate the technical skill of laparoscopic surgeons with tools like GOALS, however time for task completion is often seen as one of the most predictive objective forms of evaluating technical skill, as seen in Fig. 2.1.

Table 5.1: Likert-scale speed perception questionnaire for a manually sped-up video.

Score	**Does this video appear to have been altered to have an increased playback speed?**
)	Definitely sped-up video
)	Likely sped up video
)	Not sure / I don't know
)	Not likely sped up video
)	Definitely real-time playback speed (unaltered speed)

2.1 Biological Motion Perception

Previous research involving point light walkers (PLW), 10-12 dots of light illuminated on a screen to correspond to the human joints in walking motion, (illustrated in Fig. 2.13), revealed that humans are

Seattle Children's — Internal

excellent at recognizing more subtle qualitative characteristics like the gender or emotion of a PLW by inferring the gait from a few moving points [62]. Once the PLW's walking gait speed is changed however, the abil- ity to recognize gender is diminished [61]. This and other work suggest that biological motion perception is somehow tuned to the speed of the motion involved [2].

Due to task time being such a strong indicator of surgical technical skill, we speculate about whether a video artificially sped up to simulate a quicker time for task completion will lead to an altered perception of technical skill, or if human raters evaluate other, more nuanced, qualities of the performance.

2.2 Objective Metrics

Other metrics exist for measuring technical skill, with varying degrees of success at classifying novice and expert skill levels. These techniques utilize kinematic tool

tip data, as opposed to using video frames to make measurements. Metrics which use algorithms or computational methods to evaluate performance are sometimes referred to as Automated Performance Metrics (APMs) [41]. The simplest metric to measure, and often the most accurate, is time of task completion. The following APMs seek to compute information more complex than speed to account for technical skill, by conversely using the form of movements made.

Prior work has discovered that recovering stroke patients have a decrease in 'jerky' non-smooth movements as they progress in rehabilitation therapy [3]. The jerk cost of all tool movements performed in a task has also been used to measure accuracy and skill in areas in which skill is needed [79]. The jerk cost is computed by taking the integral of the squared jerk summed for both the left and right hand over the course of the performance, shown in Equation 5.1.2.2, in which T is total time of task performance and x is the magnitude of the movement.

$$\text{Jerk} \int_0^T |\dddot{x}(t)|^2 \, dt \tag{5.1}$$

Spectral Arc length (SAL) is also related to smoothness of a movement, and aims to use the tangential velocity of hand movements to compute the overall smoothness of a task. This has been used to differentiate skill levels of medical professionals in the past [83], [27]. To compute the SAL, the motion in a task must be segmented into specific hand grasping movements, with the corresponding speeds and durations known for each grasp. The Fourier magnitude spectrum transform of the speed profile is then computed and normalized with respect to its zero-frequency value. The smoothness is then measured for each segmented grasp, shown in Equation 5.1.2.2, in which ω is

frequency, ω_c is cutoff frequency, and V is the Fourier magnitude spectrum of speed. Finally, a weighted average is taken across all grasping motions to arrive at an SAL
metric for the task as a whole.

$$SAL = \int_0^{\omega_c} \frac{1}{\omega_c} \sqrt{\left(\frac{d\hat{V}(\omega)}{d\omega}\right)^2 + 1^2} \, d\omega$$

$$\hat{V}(\omega) = \frac{V(\omega)}{V(0)} \tag{5.2}$$

Instead of relying on the form of a grasping movement, taking the sum of all movements which occur is an additional metric that can be used to evaluate skill in some domains in which it is believed that novice performers on average perform more grasping movements than experts. A movement can be calculated in several ways, but the most common way of recording is to define a threshold of speed at which when the performer's speed falls below that threshold, a movement has ended and a new grasping motion starts once the magnitude of the velocity has risen above that threshold again. Counting each time this threshold is passed gives this total count of all movements conducted during the task [84].

All of these APMs which may be used to compute the skill of a surgeon have in common that once the speed playback of the data is manipulated, the results normally do not lead to any more separation between skill levels, as changing the magnitude of speed will not affect calculations to a large degree.

Our motivation in this work is to learn whether, by increasing video playback speed, novices and experts will be perceived by crowds as more skilled when they appear to be moving faster, *and* if experts and novices will have different rates of perceived change as the playback speed is increased. We will do this by evaluating the ability to discriminate obvious novices from experts of both human raters and popular tool motion metrics (APMs), at different playback speeds. We expect each APM's ability to discriminate skill to operate as the experimental control in this study, with a negligible change in separation between skill levels as speed is increased. Finally, we seek to investigate how an evaluator's likely conscious awareness of whether a video is sped up or slowed down relates to their perception of skill.

1.3 Methods

3.1 Dataset

This study used the Basic Laparascopic Urologic Study (BLUS) dataset, described in detail in [22]. This dataset arose from a gap in the field, in which no educational surgical certification process existed for urologic surgery, as opposed to how the Fundamentals of Laparosocopic Surgery (FLS) exists for general surgical procedures [75–77]. The BLUS training curriculum aimed to address urology appropriate skills improvement by recording video performances in an initial validation project of over 450 videos [78].

This dataset contains 454 videos of surgical performances consisting of four surgical tasks (110 peg transfer, 110 pattern cutting, 115 suturing, 119 clip applying), which are performed by medical students, urology residents, fellows and faculty surgeons from eight academic urology training centers in the United States [79]. Each trial of a surgeon performing one of the four tasks was recorded at 30 fps with a fixed camera-position of the laparoscopic tools interacting with the training field. Each trial additionally has kinematic data, sampled at 30 Hz, logging the tooltip positions, grasping force, and the jaw angles during the performance, as well as demographic information for each performer being obtained. A GOALS score for each performance was also previously obtained from either expert or crowd evaluation.

In previous research the Peg Transfer task has been shown to be one of the most easily differentiable tasks for surgical technical skill. Although the Peg Transfer task is criticized as the least clinically relevant, its clear ability to separate novice/expert skill levels was preferred to explore the research questions herein. Ten videos from the peg transfer task were used, in which five 'obvious experts' and five

Seattle Children's — Internal

'obvious novices' were chosen as baseline definitions of skill [70]. Here an obvious novice is defined as someone who should never be allowed to operate and obvious experts as surgeons who should never be disqualified from operating. An obvious expert was chosen such that the performer was in the top 20% of experience levels (attending surgeon or faculty urologist), previously obtained GOALS scores, and task completion times of all peg transfer tasks. The obvious novices were chosen in the same fashion such that they were in the bottom quintile of these domains. This method aimed to provide two well discriminated clusters of skill levels.

Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each non-expert crowd worker was paid an average of $0.10 to watch and evaluate a video, depending on the video duration. A web domain was created for which Turkers would be redirected to, where they submitted a consent form and were asked questions about videos. Two different kinds of experiments were conducted: *technical skill perception* and *speed perception*.

3.2 Experiment 1: Technical Skill Perception

Technical skill perception was measured by surveying non-expert crowds to give each video performance a GOALS score by rating each of the 4 subdomains shown in Table
2.1. Forty "turkers" were recruited per video, in which each video at each playback speed was independently submitted to the website in order to avoid a grouping bias. Videos were altered to speeds in the range 0.4x-3.6x, moving in intervals of 0.4, edited using FFmpeg [85], in which frames were either taken out or added in order to create

the resulting playback speed. The score from each of four subdomains (Depth Percep- tion, Bimanual Dexterity, Efficiency, and Tissue Handling) were summed to create a cumulative score for each performance in the range of 4-20. The mean of each video's cumulative GOALS score was recorded, with a 95% confidence interval.

3.3 Experiment 2: Speed Perception

The video playback speed perception test was conducted by using one novice and one expert from the peg transfer task, and asking the evaluator if they thought the video was being manually edited. If we showed them a video with a decreased playback speed, the worker was asked whether they thought the video was running at real-time speed or was manually edited to be slowed down. In the same way, workers were asked if the video was being sped up or played at real-time playback speed for videos which had an increased video playback speed. These Likert-scale questions are shown in Table

5.1 for a video in which the playback speed was increased. Speeds tested ranged from 0.2x-3.0x, moving in intervals of 0.2x to verify the range of speeds to use for the technical skill perception tests.

3.4 APM Validation at Each Playback Speed

The Automated performance metrics mentioned in Section **??** were also recorded at these various playback speeds for comparison to crowd evaluations. These were recalculated by scaling the magnitude of the speeds in the kinematic data to be equivalent to the speed displayed in a given video playback speed, then computing these objective metrics. For the jerk cost, this involved using a Holoborodko smooth noise-robust differentiator [86]. For the others this involved simply scaling the magnitude of speed and inputting the result to the pre-created function for each corresponding APM. All APMs were computed in Python and MATLAB, [87, 88].

1.4 Results

5.1.4.1 Technical Skill Perception

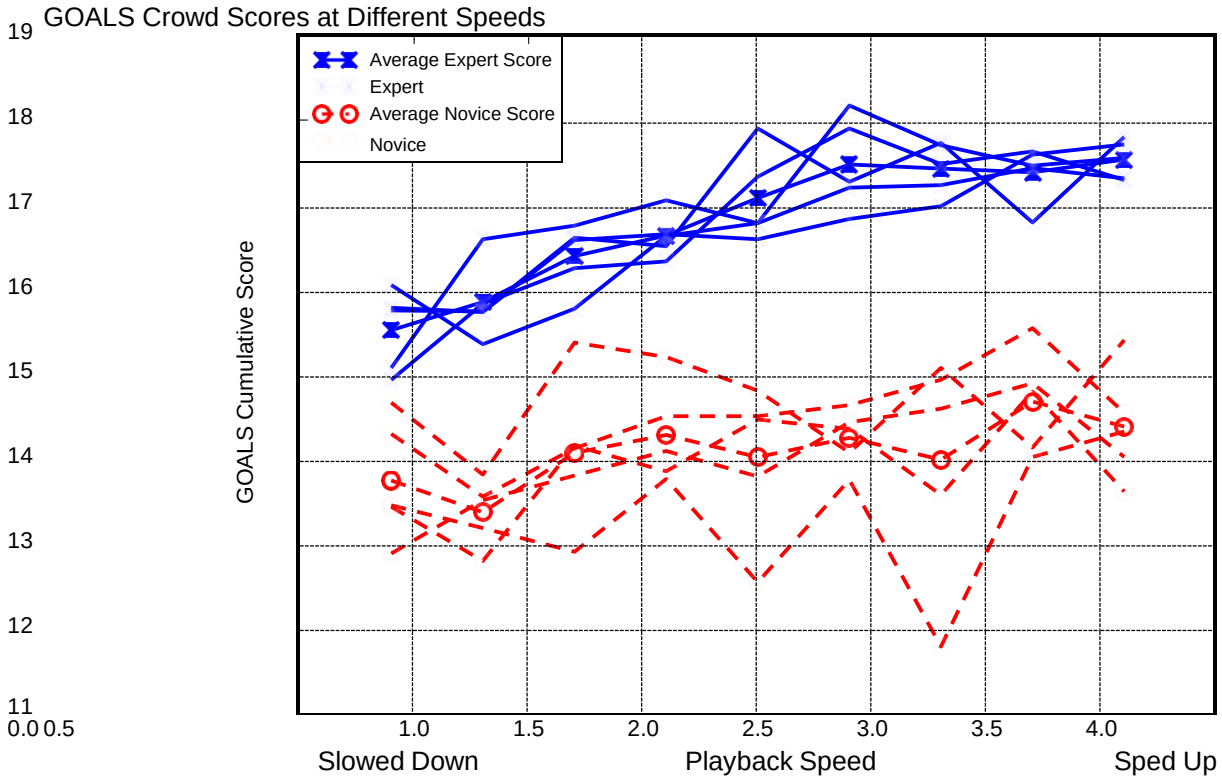
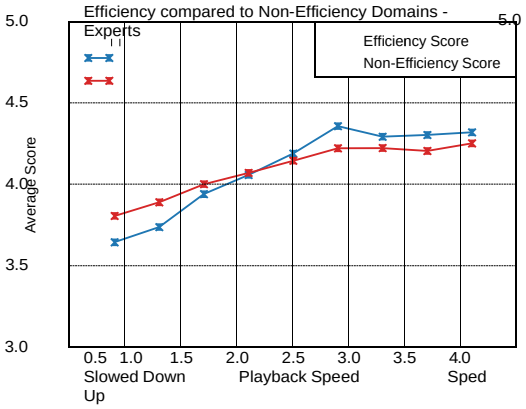


Figure 5.1: All mean crowd evaluations from each novice and expert peg transfer task at various video playback speeds. (Each solid marker indicates N = 40).

The mean of the GOALS score for each obvious expert and novice are shown in Fig.

5.11. For expert videos, each increase in the playback speed by 0.4x was associated with, on average, a 0.72-point increase in the GOALS score (95% CI: 0.60-0.84 point increase; $p < 0.001$). On average these scores appear to increase within a sublevel of the playback speeds around 0.5x to 2.4x, and then level out at all remaining playback speeds. For novice videos, each increase in the playback speed by 0.4x was associated with, on average, a 0.24-point increase in the GOALS score (95% CI: 0.13-0.36 point


Efficiency compared to Non-Efficiency Domains - Experts
Efficiency Score
Non-Efficiency Score
Average Score
5.0
4.5
4.0
3.5
3.0
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Slowed Down
Playback Speed
Sped Up

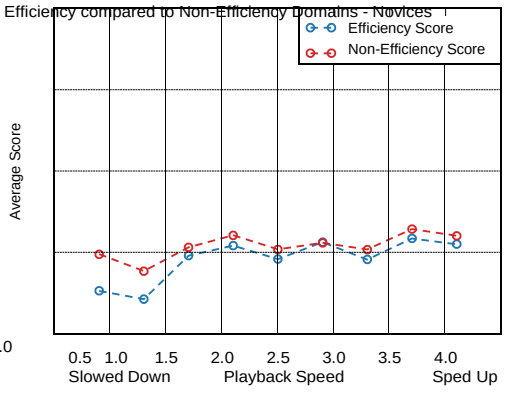
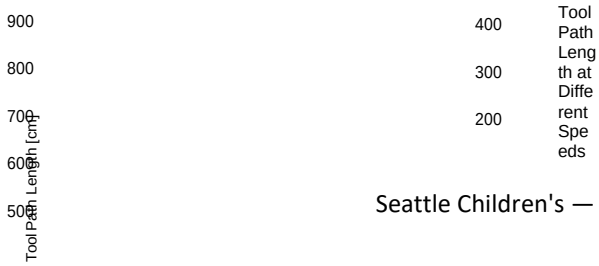

Efficiency compared to Non-Efficiency Domains - Novices
Efficiency Score
Non-Efficiency Score
Average Score
5.0
4.5
4.0
3.5
3.0
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
Slowed Down
Playback Speed
Sped Up

Figure 5.2: The efficiency subdomain as compared to the average of the other GOALS subdomains, for experts and novices.

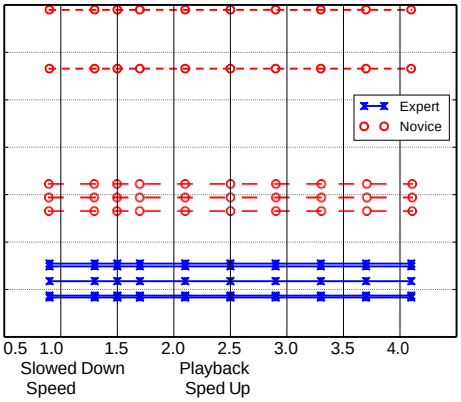
increase; $p < 0.001$). Thus, while both experts and novices had increased perceived technical skill as the playback speed was increased, the gain was significantly greater for experts. The experts had, on average, a 0.47-point greater increase (95% CI: 0.31-0.64; $p < 0.001$) in the GOALS score, compared to novices, for each 0.4x increase in playback speed. Fig. 5.12 shows the increase in the efficiency subdomain as well as the average of the other three domains to visualize whether efficiency (seen as the most related to speed) is the only increasing domain. As shown, it is clear that the other domains increased at nearly the same rate.

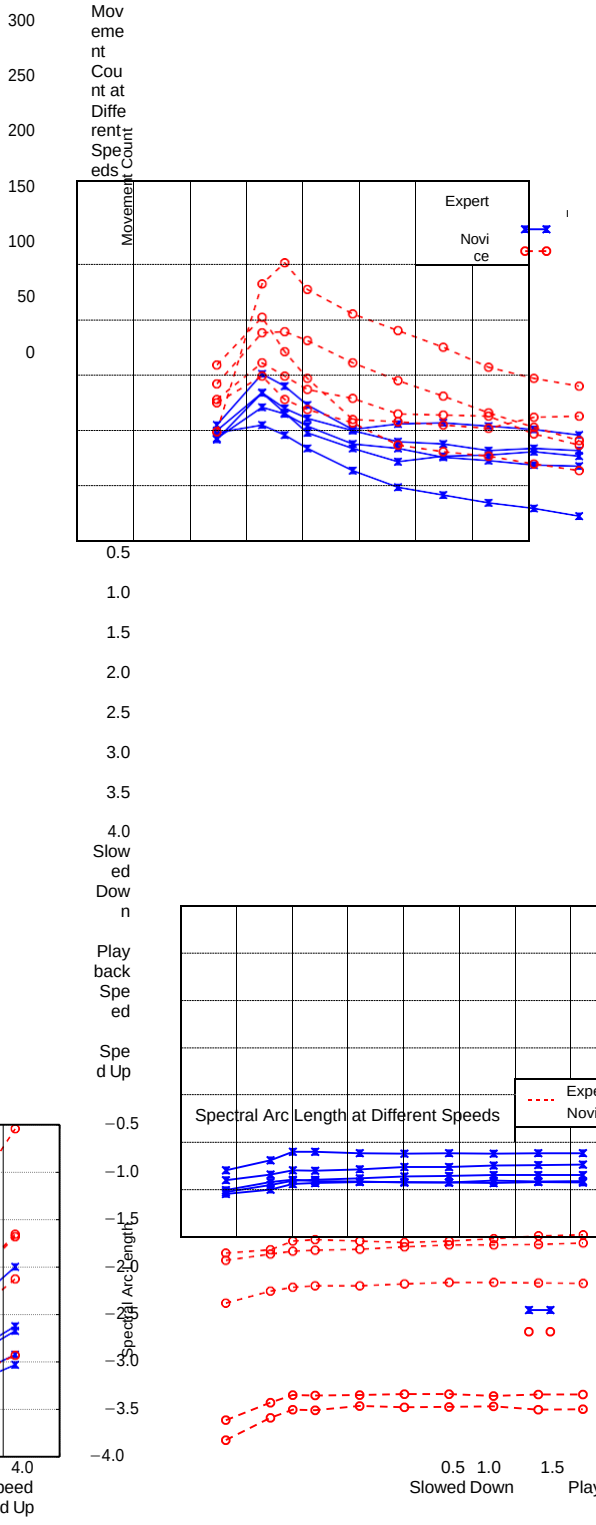
The APMs which were calculated with the manipulated kinematic data are shown in Fig. 5.3. As shown, most metrics show very little change at each of the different speeds, but even if they change, performers' technical skills are not able to be discriminated between expert and novice skill any more easily. These metrics were also calculated by incorporating a combination of scaling the speeds and removing indices corresponding to the frames which would be removed from videos when the playback speed is increased. These figures aren't included, but the results exhibit the same lack of separation between the skill groups at various playback speeds.

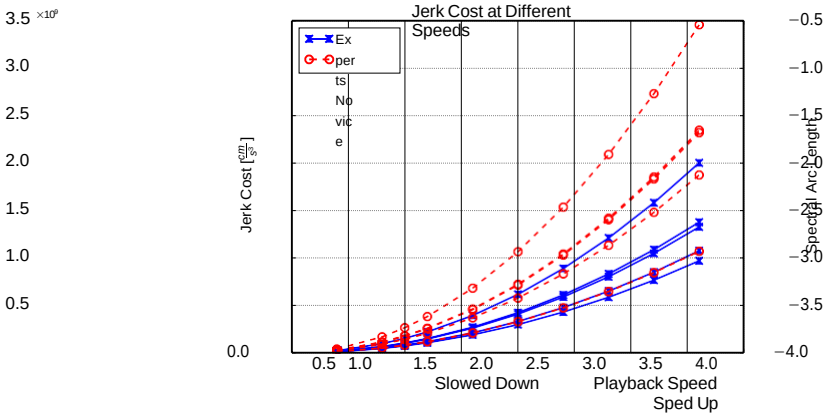
Fig. 5.4 shows the difference in mean for experts and novices, tested with both the crowd scores and the highest performing APM, spectral arc length, to illustrate the difference in skill discrimination between crowd workers and the most accurate APM as speed is increased.



Seattle Children's — Internal







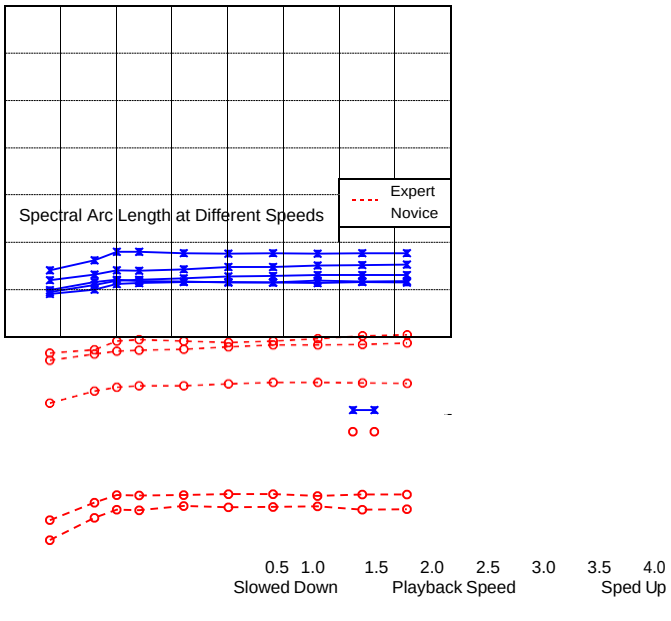


Figure 5.3: Objective technical skill metrics for the obvious experts and novices across various speeds.

5.1.4.2 Playback Speed Perception

The speed perception results for an obvious novice and obvious expert are plotted in Fig. 3.5a. The scores were obtained by treating each of the scores in Table 5.1 as a score, and obtaining the mean score for each video evaluated, such that a higher score leads to a higher perception of the video being played at a real-time playback speed. Both novices and experts displayed similar perceptions of the video either having an increased or decreased video playback speed. In Fig. 5.5b, a two sample independent t-test was computed between the two groups at each of the playback

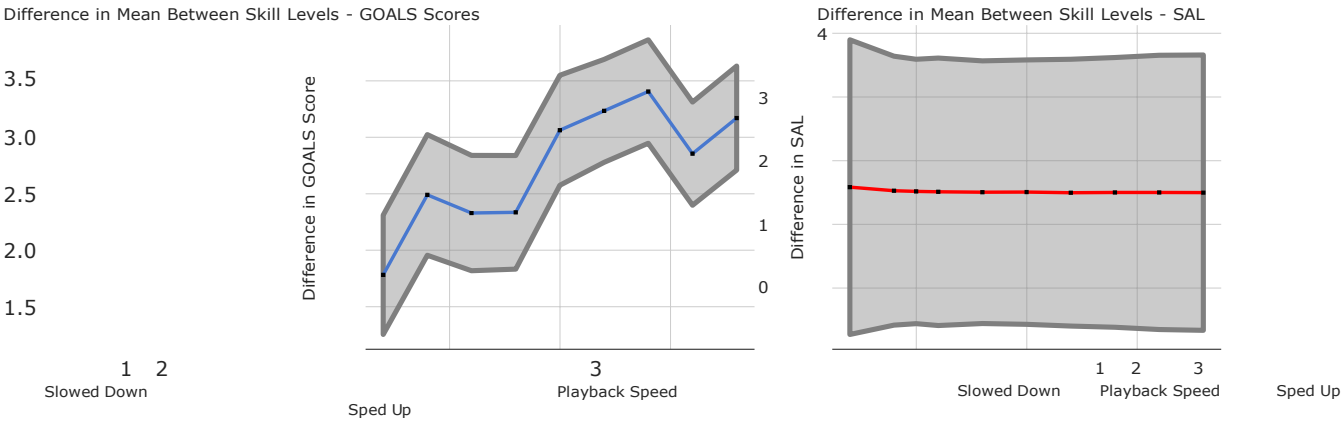
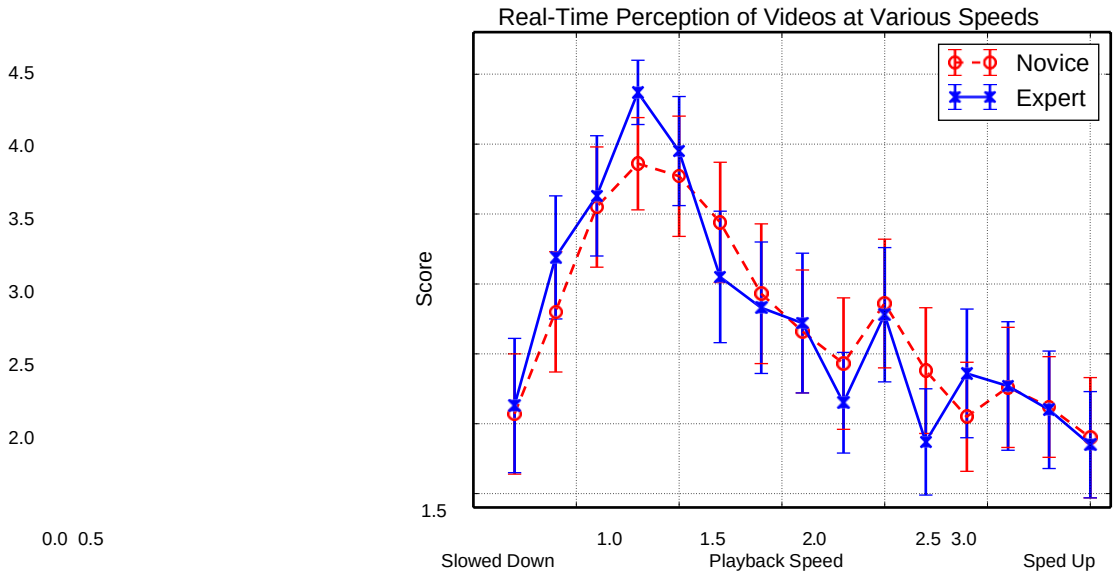


Figure 5.4: Difference in mean between skill levels from the highest performing APM and from crowd scores, with a 95% CI in the shaded regions.

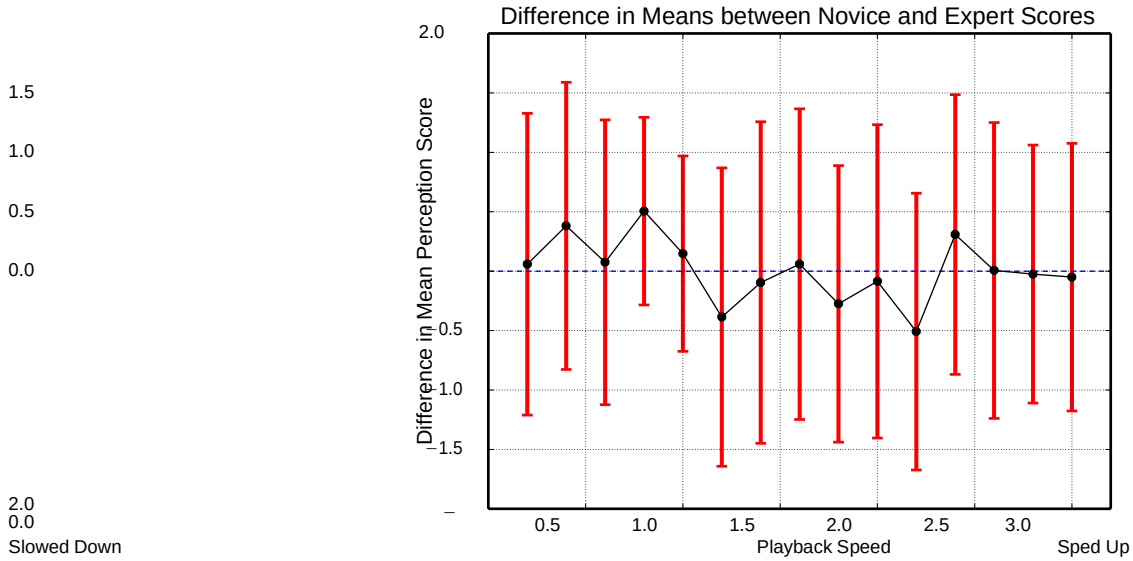
speeds, to visualize the difference in mean speed perception between experts and novices. All of the speeds show no significant difference in playback speed recognition. The perceived video playback speed scores are compared with the same video's crowd scores for the GOALS assessment metric in Fig. 5.6.

1.5 Conclusion

The results from the technical skill perception study give support to our initial hypothesis that increasing the video playback speed would increase the ratings of surgical performances. Surprisingly, however, we discovered that novice performances receive a much lower increase in score, which is almost negligible. This finding elucidates the notion that despite increasing the video playback speed of a novice performance, non- expert crowd workers are still able to spot the more obvious mistakes made by these

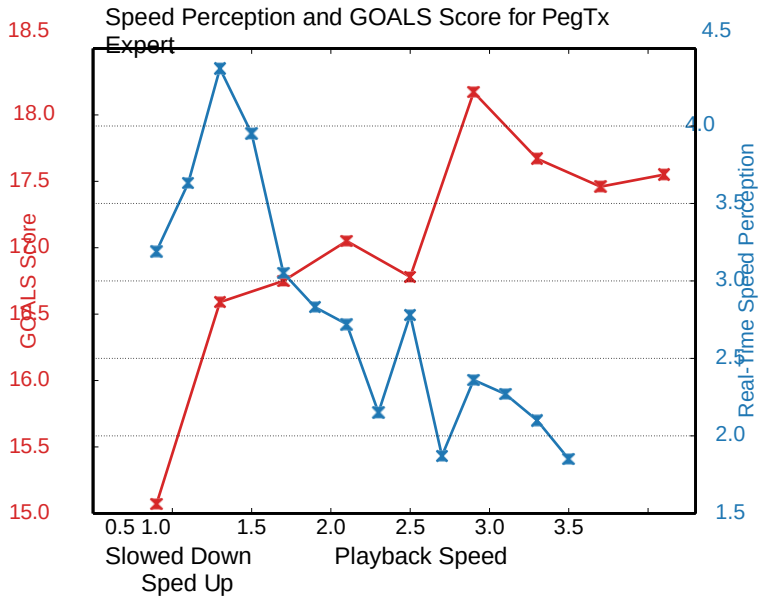


(a) All mean crowd evaluations of video speed perception at various video playback speeds. (Each solid marker indicates N = 40).

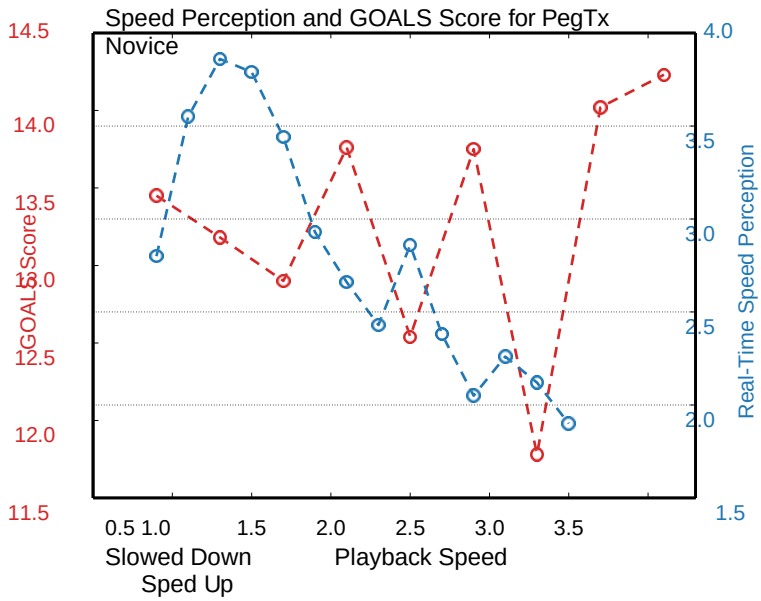


(b) Difference in Mean between Expert and Novice evaluations.

Figure 5.5: Comparison of Novice and Expert Speed Perception at each playback speed. (There were N = 40 *unique human evaluations/skill level/playback speed.* The error bars represent 95% confidence intervals.)



(a) Expert



(b) Novice

Figure 5.6: Speed Perception compared with technical skill perceived at various speeds.

novices. However, it appears that crowd evaluators are biased to speed when they eval- uate expert performances. This could be due to expert performers appearing as though

their movements are even smoother at quicker playback speeds and the few mistakes they are making being 'washed out' or not emphasized at quicker playback speeds. With 135,000 surgeons in the U.S. and the growing need to objectively quantify surgical skills to ensure public safety, methods to rapidly triage technical skill are required. Our obser- vations may provide an easier and cheaper way of discriminating novices from experts than with using expert surgeon evaluation.

To illustrate whether the 'Efficiency' domain of the GOALS assessments was the only domain increasing, as this domain is most closely related to speed, a scatter plot for the average efficiency domain and average of the other three domains for novices and experts is shown in Fig. 5.12. The novices clearly show no difference in Efficiency vs. the other three domains. In addition, the three remaining domains for experts appear to increase at almost the same rate as the Efficiency domain. It appears that crowds are also biased to give higher ratings in other domains which aren't necessarily related to speed.

The results from the real-time playback perception study validate our initial hypoth- esis by showing us that crowds are able to accurately discern when a video has been manually edited to have a different playback speed. Although, as shown in Fig. 5.5a, there is apparently no difference in perception between an obvious novice performance and an obvious expert performance.

As shown in our comparison to objective metrics, no other used method of objectively obtaining technical skills scores will show the same amount of improvement for increases in playback speed. There may be information that humans can decipher which objective metrics or machine learning algorithms cannot.

We conclude that increasing the video playback speed of performances of dry lab laparoscopic training tasks could provide a cheap and easy way to discriminate experts from novices as the separation in GOALS scores between novices and experts appears

to increase as video playback is increased. A limitation of this study is that we only sampled ten videos of laparoscopic training procedures. Additional investigation with a larger dataset of more clinically relevant performances is required to conclude whether this observation extends to actual surgical case evaluation.

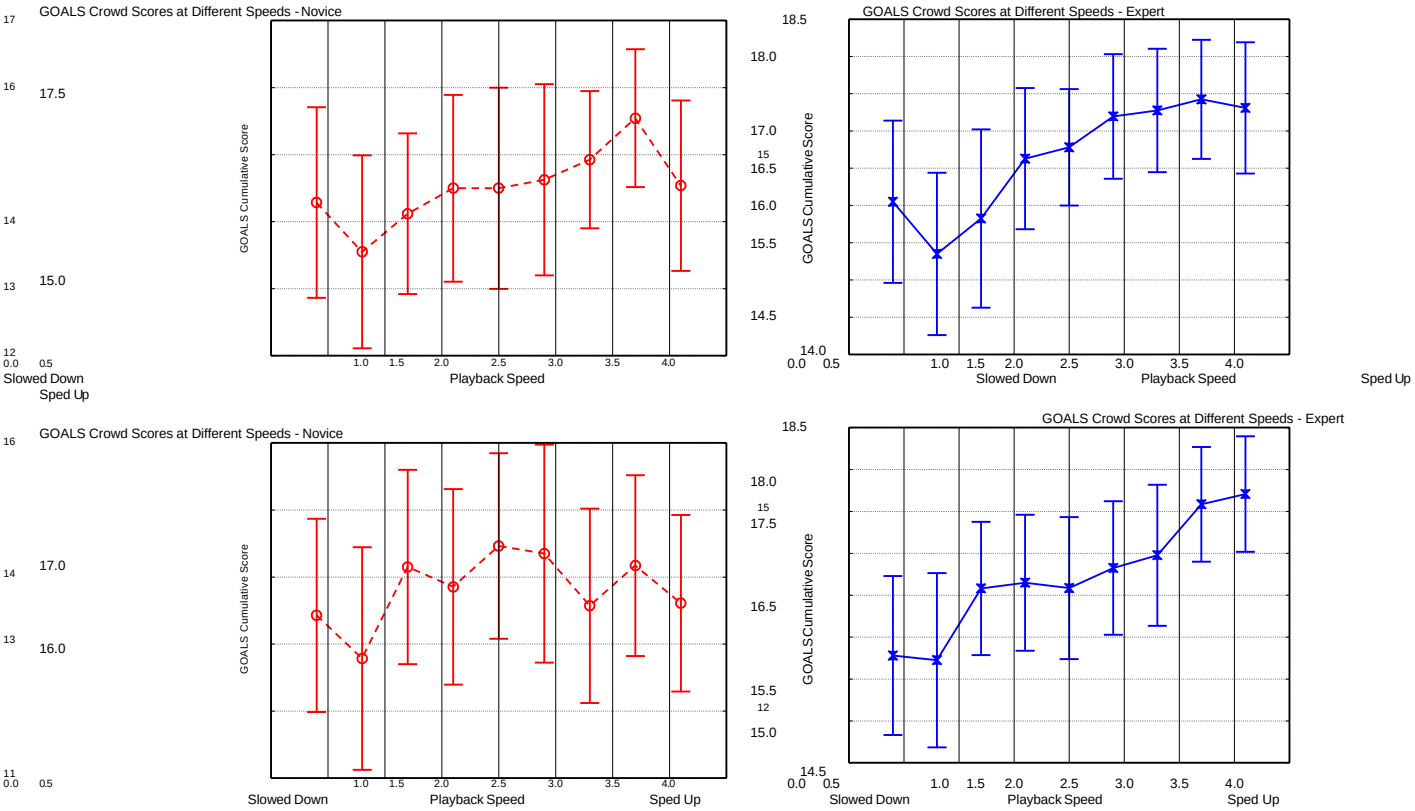


Figure 5.7: Crowd GOALS evaluations for Experts and Novices at various playback speeds. Each point is N = 40, with 3600 total evaluations.

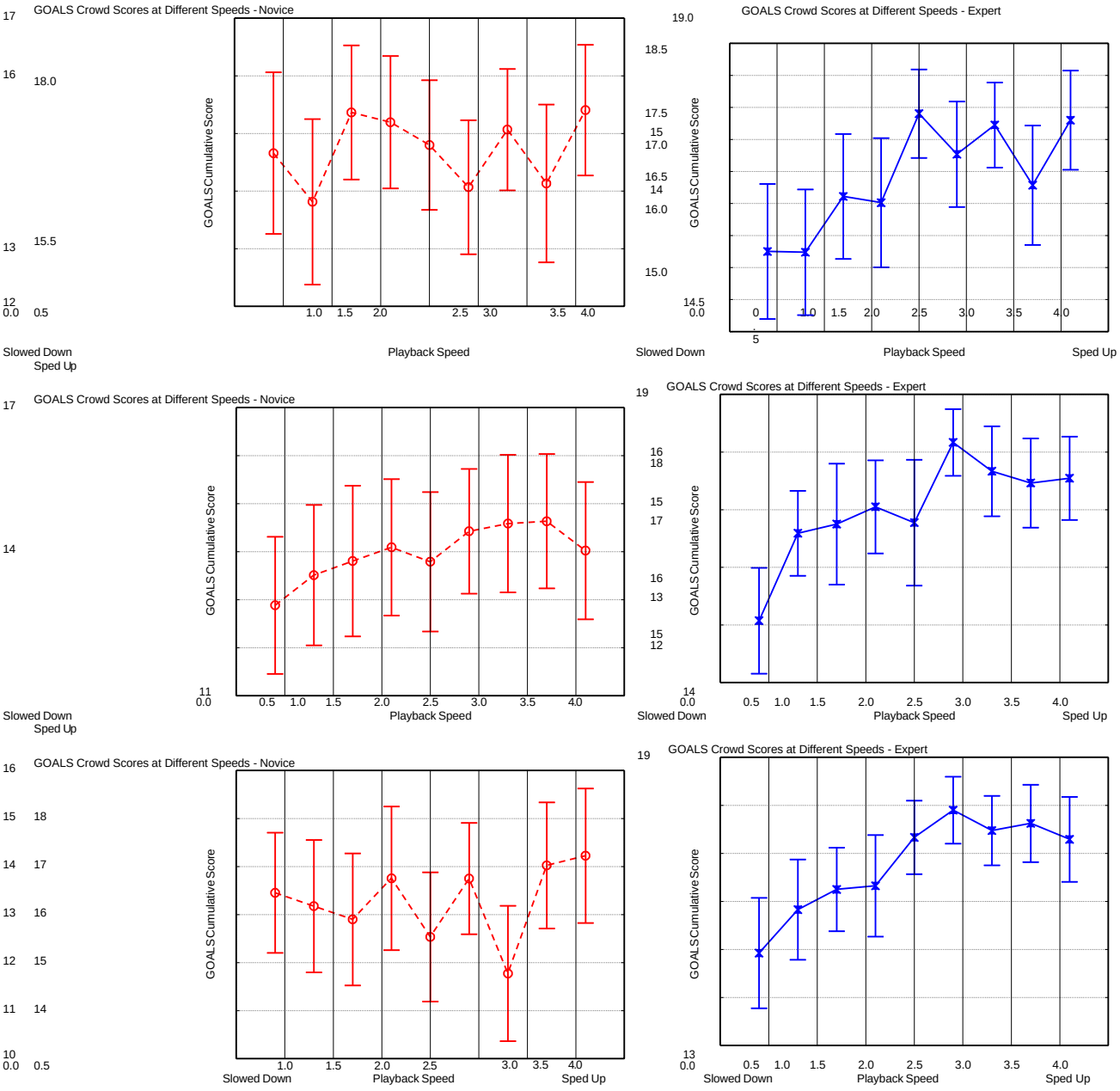


Figure 5.8: Crowd GOALS evaluations for Experts and Novices at various playback speeds. Each point is N = 40, with 3600 total evaluations.

5.2 The Effect of Video Playback Speed on Perception of Technical Skill in Robotic Surgery [81]

The previous study had two major shortcomings, it was a small amount of data and the videos being evaluated weren't from actual surgical footage. This following study aimed to dive deeper into the notion of speed perception in surgical settings and learn the effects of studying a smaller subset of skill, surgeons who are all above average per- formers.

2.1 Abstract

Purpose: Finding effective methods to evaluate surgeon technical skill has proven a complex problem to solve computationally. Previous research has shown that obtaining non-expert crowd evaluations of surgical performances concords with the gold standard of expert surgeon review, and that faster playback speed increases ratings for videos of higher-skilled surgeons in laparoscopic simulation [22], [73]. The aim of this research is to extend this investigation to real surgeries that use non-expert crowd evaluations. We address two questions (1) whether crowds award more favorable ratings to videos shown at increased playback speeds, and (2) if crowd evaluations of the first minute of a surgical procedure differ from crowd evaluations of the entire performance.

Methods: A set of 56 videos of practicing (non-novice) surgeons including robotic prostatectomy, hysterectomy, and partial nephrectomy (for 28 "expert" surgeons, and 28 who are "proficient"), were used to evaluate the perceived technical skill of the surgeons at each video playback speed used (0.4x, 1.2x, 2.0x, 2.8x, and 3.6x) for the first minute of the previously rated performance, using the Global Evaluative Assessment of Robotic Skills (GEARS) assessment criteria. Each video was subsequently rated at 1x speed to obtain objective ratings for the first minute of the surgical procedure. **Results:** Crowds on average did rate videos higher as playback speed was increased. This effect was observed for both proficient and expert surgeons. Each increase in the playback speed by 0.8x was associated with, on average, a 0.16-point increase in the GEARS score for expert surgeons and a 0.27-point increase in GEARS score for proficient surgeons with both groups being perceived as obtaining relatively equal skill at the fastest playback speed. It was also found that 22 out of the 56 surgeons were perceived to be significantly different in skill when just viewing the first minute of performance, with 11 of the 28 surgeons in both skill categories being rated as belonging to the opposing category.

Conclusion: The observed increase in skill ratings with video playback speed replicates findings for laparoscopic experts in [73], and extends to the context of real robotic surgeries. The change in perceived technical skill due to increased playback speed for experts and proficient surgeons suggests that crowds do seem biased in rating surgeons as more highly skilled when they appear quicker, even if speeds seems unrealistic. Furthermore, the large differences in skill labels when comparing the first minute of surgery to the entire 15 minute video warrants further investigation into how much perceived skill ratings vary in time (sub-task level) vs. summative metrics (task level).

2.2 Introduction

A third of all deaths in the United States are caused by medical errors, and sur- gical errors are one of the largest contributors to this [4]. Technical surgical skill is directly related to patient outcomes [72], but it remains a difficult computational task to correctly classify surgeons into skill levels with a compelling level of accuracy, i.e. never misclassifying an 'obvious novice' as an 'obvious expert' and vice versa - the MAC criterion [70]. The de facto gold standard for evaluating technical skill is video evaluation by an expert surgeon using Likert-scale assessment metrics, in which evaluators submit ratings on an anchored scale of 1-5. Using crowds of *non-expert* evaluators are a surprisingly accurate way to inexpensively and rapidly obtain skill level ratings for videos of surgical performances, with a pass/fail rating that matches 100% of pass/fail

ratings by expert surgeons [22]. The fact remains, however, that humans can be biased in their thinking, and subjective metrics of rating performances can lead to results we would not expect from computational models of evaluation.

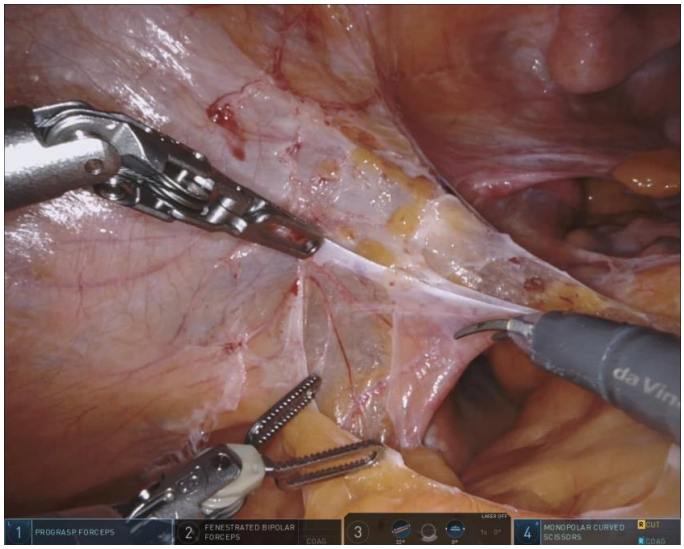


Figure 5.9: Frame from a video collected for this study, recorded from a daVinci surgical robot, Intuitive (Sunnyvale, CA)..

A popular laparoscopic surgical skill assessment metric is the Global Evaluative Assessment of Robotic Skills (GEARS), which is the most common assessment tool for robotic surgery skills [89]. The subdomains in this metric include: bimanual dexterity, efficiency, depth perception, force sensitivity and robotic control. Task time is not a direct metric used to evaluate the technical skill of laparoscopic surgeons with tools like GEARS, however time for task completion is often seen as one of the most predictive objective forms of evaluating technical skill [13], [41]. However, there must be ways of objectively evaluating skill between multiple performances which were completed in the same time span.

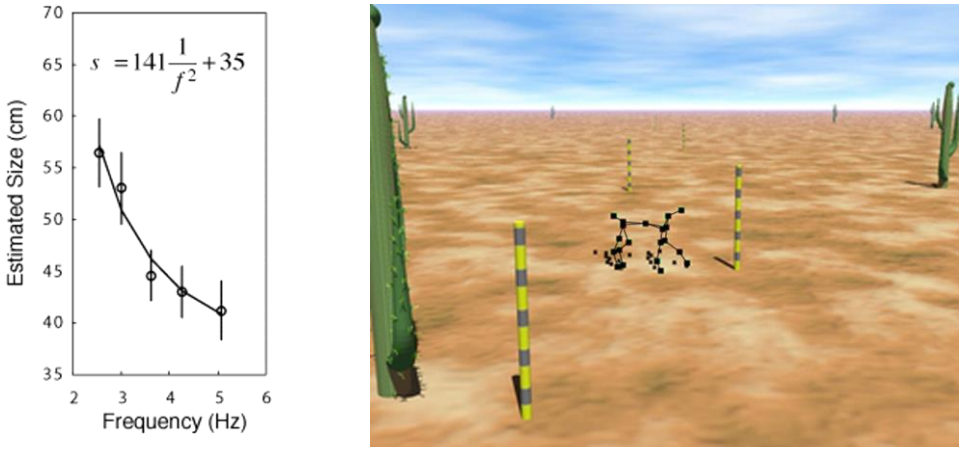


Figure 5.10: Illustration of research results (left) which found the frequency of gait in simulated animals (right) affects the estimated size [**?**].

2.1 Changes in Perception due to Speed

Research in biological motion processing has observed a relationship between the gait frequency and estimated size of animals and other objects which replicate the joints of the human body. It was found that when a computer simulation was used to artificially modify the gait speed of these objects to unnaturally high speeds, study participants perceived the size of the moving objects as changing. As the gait frequency was increased, the estimated size of the object decreased, shown in Fig. 5.10, [66], [61]. The results from this study may suggest a link between how humans evaluate biological motion and the speed of a movement. It is possible this phenomenon extends to other areas of evaluation, such as technical skill in surgery.

2.2 Technical Skill Ranges

Previous work has shown that crowds are able to discriminate the levels of novice and expert surgeons and their scores change as playback speed changes - even into motions appearing artificially sped up [73]. This was a limited study using a small dataset of laparoscopic training exercises from the Basic Laparoscopic Urologic Skills dataset [78], [79], with extremes of skill levels (e.g. “obvious novice” medical students with no suturing experience). To test whether this phenomenon relates to actual surgical procedures, a larger dataset must be used which incorporates videos of real surgical footage and more practical ranges of surgeon skill. However, surgeons which are allowed to safely operate on patients can’t ethically be “obvious” novice surgeons, according to the MAC Criterion [70]. This means the surgeons in a dataset of real surgical footage will comprise a smaller range of technical skill levels, with all surgeons being more comparably rated. This necessitates the need to further separate experts into two groups: proficient and expert surgeons. Proficient surgeons as stated in [90] are defined as a surgeon who is well advanced in any branch of knowledge or skill. We will adopt this as our term for surgeons who are in the bottom quintile of scores from our surgical videos, as they are still well above average skilled surgeons.

Seattle Children's — Internal

Our motivation in this work is to investigate whether, by increasing video playback speed in real robotic surgery, surgeons will be perceived by crowds as more skilled when they appear to be moving faster, *and* how these effects vary from expert and proficient surgeons. We will do this by evaluating the ability of non-expert crowd workers to discriminate proficient surgeons from experts. Finally we seek to examine the effect of video duration on ratings following [91].

2.3 Methods

3.1 Dataset

This study used a novel Robotic Surgery Readiness (RSR) Study dataset, which con- sists of 343 videos of live robotic surgeries, with matching kinematic data also recorded, though unused here. These surgeries were performed by attending surgeons and trainees in urology, gynecology, and general surgery at the University of Washington Medical Center and the Puget Sound Veterans Administration. Each video was manually edited to include roughly the first 15 minutes of surgical activity performed by the surgeon. There are a wide variety of surgical procedures recorded from this dataset including: prostatectomy, cystectomy, hysterectomy, partial nephrectomy, and sacrocolpopexy. An image of a frame from one of these surgeries is shown in Fig. 5.9. A GEARS score for each performance was also obtained from crowd evaluation.

The range of scores for the RSR videos was fairly small, with most lying between 20-22 out of 25. In an effort to get the largest possible range of skill from this dataset, 28 performances from the top quintile of scored performances and 28 from the bottom quintile of performances were used and given labels of ‘expert’ and ‘proficient’, respec- tively, keeping in mind that almost all of the performances would have objectively been considered expert-like. For semantic analysis, the first minute of surgical activity was extracted from each 12-15 minute video for analysis by crowds.

Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each non-expert crowd worker was paid an average of $0.40 to watch and evaluate a video. A web domain was created for which Turkers would be redirected to, where they submitted a consent form and were asked questions about videos. Two different kinds of experiments were conducted: *technical skill perception at different playback speeds* and *sub-task level skill labeling.*

3.2 Technical Skills Perception at Different Playback Speeds

Technical skill perception was measured by surveying non-expert crowds to give each video performance a GEARS score by rating each of the 5 subdomains shown in Table 2.2. Forty “turkers” were recruited independently for each video, in which each video at each playback speed was independently submitted to the website in order to avoid a grouping bias. Videos were altered to play at 0.4x, 1.2x, 2.0x, 2.8x, and 3.6x, edited using FFmpeg [85], in which frames were either taken out or added in order to create the resulting playback speed. The score from each of five subdomains (Depth Perception, Bimanual Dexterity, Efficiency, Force Sensitivity, and Robotic Control) were summed to create a cumulative score for each performance in the range of 5-25.

Seattle Children's — Internal

A linear mixed effects model was used to analyze the significance of the various speeds to the evaluations

received. The significance of the time spent on reviewing a video and the labels given to the surgeons were also examined. It was hypothesized that each Mechanical Turker on the site should be assumed to have a different slope, to match the difference with which they relatively evaluate different videos. The evaluator ID given to each rater was assigned a random effect. In addition, fixed effects for both the speed at which the video was played and the amount of time spent on the video by the evaluator were controlled for in

the model. The mixed model was compared with a null model, which did not include the fixed effects, using ANOVA hypothesis testing and comparing the relative information criteria and correlation between the scores and the various parameters. All data aggregation and statistics were calculated in Python 3.6 [87] and R [92].

3.3 Sub-Task Level Skill Labeling

To learn if the first minute of surgical activity displayed the same level of technical skill as the entire 15 minute video and should have the same label, the first minute of each video at 1x (normal playback speed) was submitted to crowds for review. The mean of these new scores were then calculated, and the 50th percentile was used as the cut-off point to signify in this dataset of 56 videos, whether a surgeon would be labeled as a proficient or expert level surgeon. The differences in these labels were then analyzed further and visualized, to find if the previously obtained scores for the entire 15 minute video possessed significantly different skill levels in the first minute, such that the label of the corresponding surgeon was different than what we had originally obtained.

2.4 Results

Table 5.2: Results from the linear mixed effects model testing speed and time spent on evaluation

Fixed Effects	Estimate	Standard Error	df	t value	Pr(>|t|)
Initial linear mixed effects model (BIC = 52465.36)					
Speed	0.046	0.013	13360	3.47	5.31e-04
Elapsed Time	-0.0023	3.47e-4	13720	-6.57	5.18e-11
Final linear mixed effects model (BIC = 52410.60)					

2.5 Data Demographics

Table 5.3 summarizes the types of robotic surgery and the demographic data from the $N = 56$ videos (and surgeons) used in this work.

Table 5.3: The types of robotic surgery performed and demographics of the surgeons from the 56 videos used.

Surgery Type	N
Prostatectomy	17
Hysterectomy	11
Nephrectomy	5
Cystectomy	4
Other	19

Demographic	N
Male	17
Female	11
Mean Age	46.70
Mean Yrs. Experience	12.85

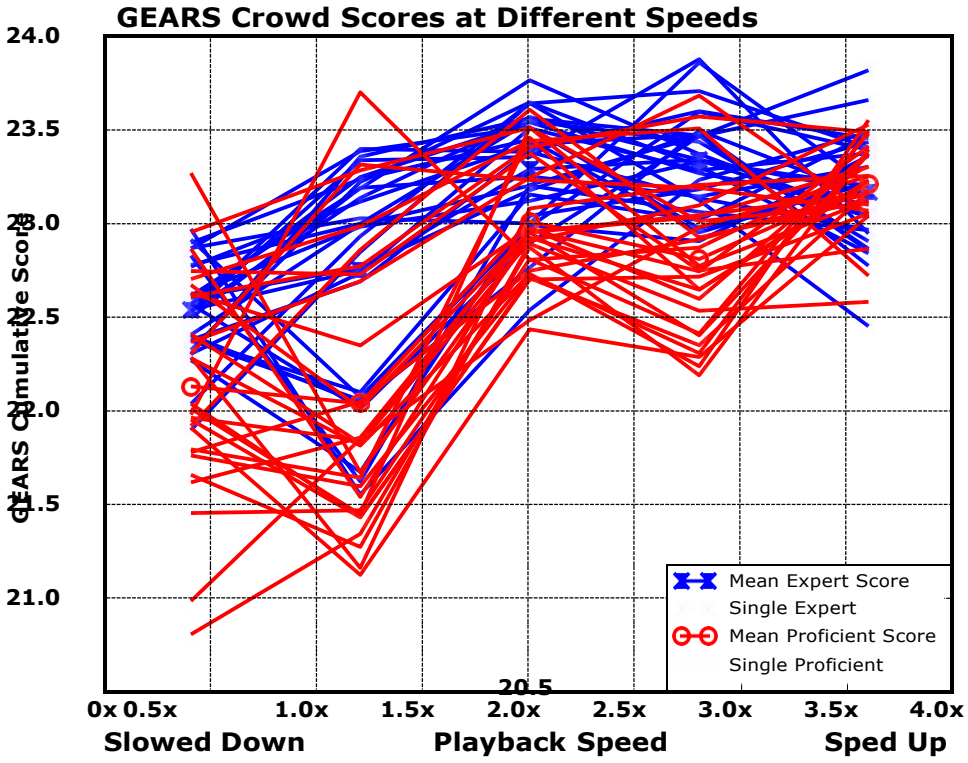


Figure 5.11: All mean crowd evaluations (bold lines) from each proficient and expert surgeon at various video playback speeds. Each single surgeons video (semi-transparent colors) indicates ratings from N = 40 turkers.

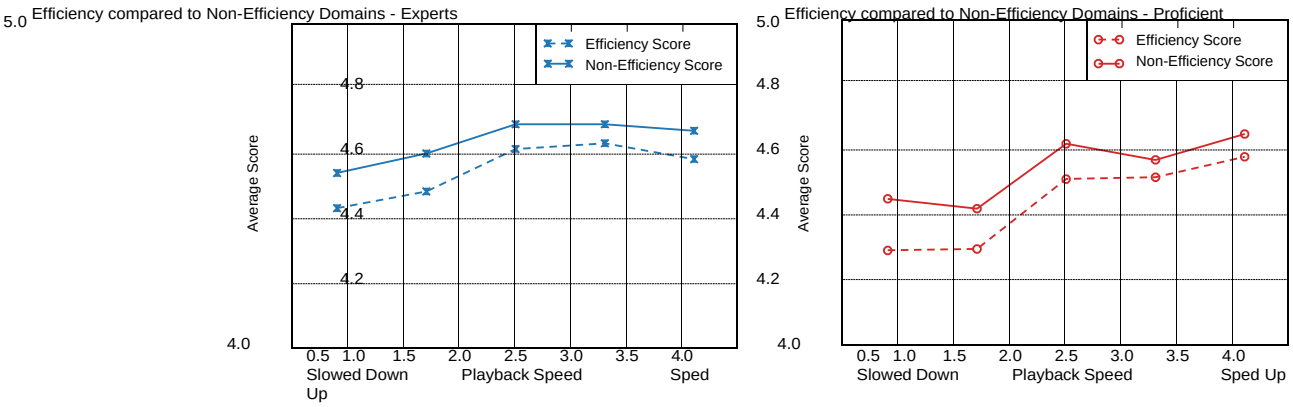


Figure 5.12: The efficiency subdomain as compared to the mean of the other GEARS subdomains, for expert and proficient surgeons.

5.1 Technical Skill Perception

The mean of the GEARS score for each group and video are shown in Fig. 5.11. For expert videos, each increase in the playback speed by 0.8x was associated with, on average, a 0.16-point increase in the GEARS score (95% CI: 0.10-0.22 point increase; p

< 0.05). On average these scores appear to increase within a sublevel of the playback speeds around 0.4x to 2.0x, and then level out at all remaining playback speeds. For proficient surgeon videos, each increase in the playback speed by 0.8x was associated with, on average, a 0.27-point increase in the GEARS score (95% CI: 0.19-0.35 point increase; $p < 0.05$). Thus, while both experts and proficient surgeons experienced increased perceived technical skill as the playback speed was increased, the mean score obtained at the fastest video playback speed reached an almost equal skill level for both proficient and expert level performances. Fig. 5.12 shows the increase in the efficiency subdomain as well as the mean of the other four domains to visualize whether efficiency (seen as the most related to speed) is the only increasing domain. As shown, no major difference is apparent between the two types of domains.

5.2 Sub-Task Level Labeling

Figure 5.13 illustrates the difference in GEARS scores by non-expert crowd workers, when viewing the first minute of the video compared to the entire video. As shown, there is a noticeable difference in the scores given at these two levels. Most surgeons, when viewed for the entire 15 minutes, are rated slightly higher than when only the first minute of surgical activity is performed. Viewing just the label given to the performance (proficient or expert), by analyzing whether the video was above or below the median score in the group of 56 videos, a total of 11 previously labeled proficient level surgeons, and 11 previously labeled expert surgeons switched the label they were originally given, when only the first minute of surgical activity was evaluated.

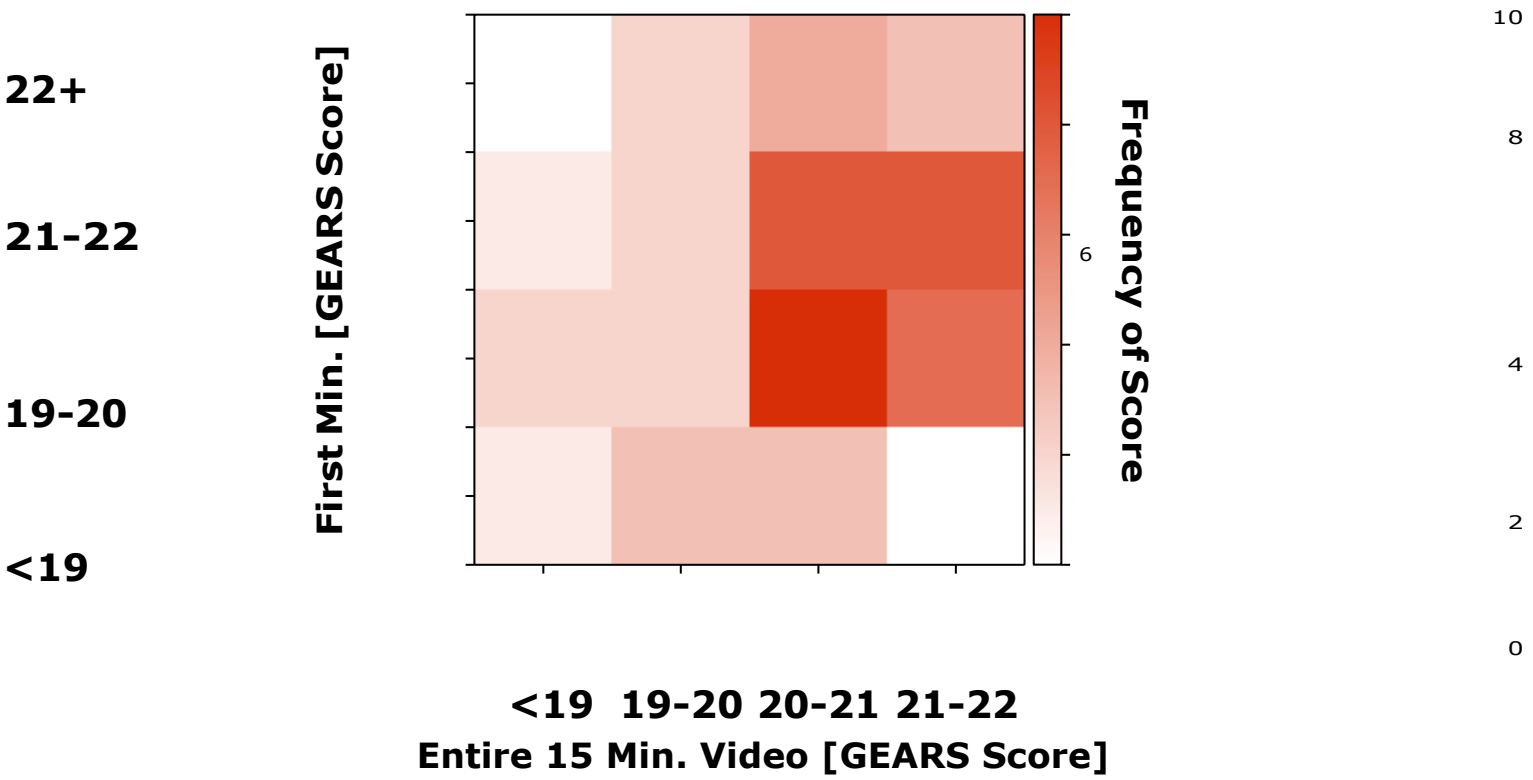


Figure 5.13: GEARS scores given to the entire 15 minute video of a performance, compared to only the first minute of the same video.

2.6 Conclusion

The results from the technical skill perception study give support to our initial hypothesis that increasing the video playback speed would increase the ratings of ex- perienced surgeons. Now this evidence extends to real surgeries using robotic surgical performances. Surprisingly, however, we discovered that for sub-groups of expert level surgeons (proficient and expert) the increase in score happens at a quicker rate for proficient surgeons than for expert level surgeons. Increasing the playback speed of a slightly less than expert-level surgeon tends to "wash out" the minor mistakes they have made in the performance, effectively making the two groups appear more equally skilled at higher playback speeds. Additionally, from analyzing the 'Efficiency' subdomain from the GEARS assessment, it appears that - surprisingly - crowds are also biased to give higher ratings in domains which are not associated with speed.

The results from the sub-task level skill labeling experiment show us that it may be necessary to have video evaluated in smaller duration segments, due to the notable disagreement with the skill level assigned to the entire video vs. just using the first minute. This lends support to the notion that a surgeon's technical skill may fluctuate in time (on the order of minutes) throughout a surgical procedure. This warrants further study of non-constant skill mapped to a single summative rating, as this would induce substantial label noise of computational skill evaluation using machine learning.

We conclude that increasing the video playback speed of performances of practicing surgeons in typical robotic surgeries results in increased scores as reported by non-expert crowds. This effect is surprisingly uniform across GEARS subdomains, even those which should be unaffected by speed. We further conclude that more studies should be done to investigate variance in time of sub-task level videos of surgical procedures, as the technical skill of a surgeon may fluctuate on the order of minutes or less during the procedure.

Seattle Children's — Internal

Chapter 6

Robotic Surgery Readiness Study

This chapter introduces the RSR study and dataset, detailing the motivation, methods, and results from this study, which contains a reproduction of a journal article to be submitted to the journal of the American College of Surgeons entitled "Robotic Surgery Readiness: A Randomized, Controlled Study of Virtual Reality Warm-Up Prior to Robot-Assisted Surgery". In addition this chapter includes a reproduction of a journal article entitled "Temporal Variability in Surgical Technical Skill Evaluation" which used a subset of the RSR data and was submitted to the International Journal of Computer Assisted Radiology and Surgery in April 2020.

Seattle Children's — Internal

.1 Robotic Surgery Readiness: A Randomized Controlled Study of Virtual Reality Warm-Up Prior to Robot- Assisted Surgery [93]

1.1 Abstract

Background: Surgical errors and complications have been linked to the volume of cases a surgeon experiences, and research has identified a distinct decrement of skill over intervals of inactivity. High stakes professions including athletics, aviation, and music have incorporated brief pre-procedural warm-ups to counter skills decay and elevate readiness. The goals of this research were to identify the optimal virtual reality (VR) warm-up module to prime a surgeon for robotic surgery, and validate it in the operat- ing room (OR) using video review and objective performance metrics in a randomized controlled setting.

Study Design: Surgical trainees and faculty at three medical centers were recruited for participation to complete VR training on proficiency tasks. Each surgeon was ran- domized to be in one of five groups. Members of each group performed six trial sessions of warm-up modules assigned to their group to prime them for surgery. The warm-up curriculum was chosen from the group with the best task time (TT) and Global Evalu- ative Assessment of Robotic Skills (GEARS) scores, obtained by video review. Robotic surgeons were randomized to either receive or not receive the chosen warm-up module before beginning surgery. Video of the first 15 minutes of the surgeons performance was captured in the OR. The GEARS assessment tool was used to review these videos and primary metrics were calculated to assess the validity of the curriculum, by both expert raters and non-expert crowd workers. Both a linear mixed effects model with a random intercept for each surgeon and a nonparametric modified Friedman test were used to

analyze the effect of warm-up.

Results: Forty-four surgeons participated in the study to find the optimal warm-up curriculum. Group 1, a lone Running Suture module, was on average 31.3 seconds faster, and had the highest GEARS scores with the least variation. The lone Running Suture module was thus chosen as the optimal RSR curriculum. Thirty-four surgeons completed 432 surgeries of which 343 videos with corresponding tooltip instrument po- sition data were recorded, making this the largest dataset of its kind. No statistically significant difference between the performance of the control and intervention groups was found. Further, no significant differences between treatment arms were found in any of the six kinematic and event outcomes used. Kinematic and event outcomes were not significantly associated with GEARS scores as quantified by repeated measures cor- relation.
Conclusion: It appears that in surgical settings, performing a simulated VR warm-up procedure before undergoing robotic surgery does not have a significant effect on the surgical technical skills of robotic surgeons. This dataset is the largest of its kind, and will be beneficial to future research efforts involving surgical technical skill.

1.2 Introduction

The Institute of Medicine reported that medical errors account for as many as 98,000 deaths each year, atop a higher number of patient complications [5]. An estimated one third of these errors are surgical and the average American can expect to have 7 surgeries in their lifetime [94–96]. Although surgical simulations role in off-line teaching is accepted [97], simulation may also have a role in identifying decrements in surgical performance and offer a system for mitigating skill decay for direct patient care [98–101]. Perez et al. expanded on this premise and challenged researchers to not only focus on psychomotor skills, but also on cognitive and perceptual skills. They concluded that skill may decay at different degrees depending on the type of skill, and that training to mitigate decay may require targeted curricula based on the specific decay signatures [99].

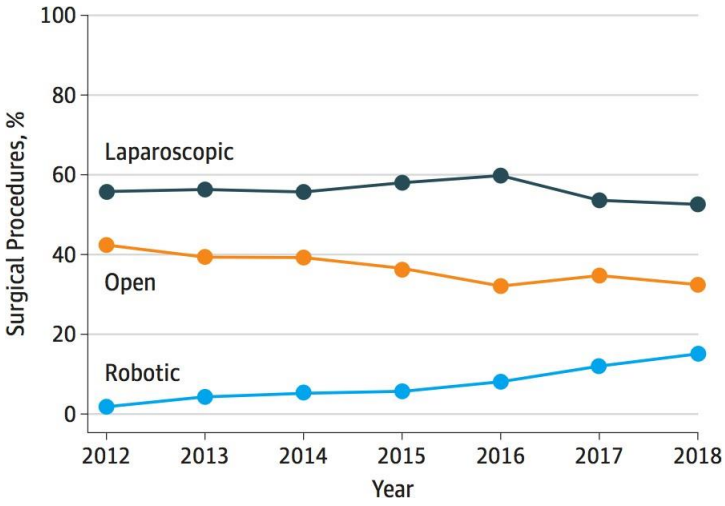


Figure 6.1: The popularity in robotic surgery has experienced a significant increase in recent years.

These methods were described for laparoscopic surgery skills, yet no intervention has been studied for robotic surgeons or reintegrated military personnel with robotic surgery practices.

Robotic surgery has experienced considerable adoption in hospitals across the United States with the creation of the da Vinci surgical robot by Intuitive Surgical (Sunnyvale, CA) [100] (Fig. 6.1). With this new technology comes the need to apply innovative methods to elevate surgeon performance.

While several other checklist-type preparation procedures exist in the OR, there is still no metric which measures how prepared a surgeon is following periods of inactivity [101]. It would be beneficial to quantify a measure of readiness to ensure a surgeon will perform at the optimal level of proficiency. It has been shown in other domains that after participating in a warm-up, task performance can be elevated [102]. Previous work additionally shows that a warm-up decrement (WUD) exists, in which the effects of warm-up begin decaying approximately 30 minutes after performing a warm-up task. [102–105]. These hypotheses have already found four factors which most affect WUD:



Figure 6.2: The da Vinci Skill Simulator (dVSS) used for surgical simulation modules, created by Mimic Technologies.

(1) intervals between practice, (2) complexity and type of task, (3) strategies employed to maintain skills, and (4) individual differences [106, 107]. Using VR proficiency-based warm-up modules have been shown to most effectively combat skill decay [108–112].

The da Vinci Skills Simulator (dVSS), created by Intuitive Surgical (Sunnyvale, CA), replaces the physical robotic arms of a da Vinci robot with virtual representations of the tools to simulate a surgical environment (Fig. 6.2) [113]. The simulation software was created by Mimic Technologies Inc. in 2007, and allows surgeons for the first time to practice surgery without any of the danger involved with real surgery (Mimic Technologies Inc). There are over 30 exercises available for practice with each of the da Vinci S, da Vinci Si, and da Vinci Xi. These tasks include suturing a wound and a typical ring-and-rail task for practicing dexterity. Measuring surgical technical skill in an objective manner is exceedingly difficult, with the current gold standard being video review by expert surgeons. Previous work has shown that video review of surgeons by non-expert crowds is surprisingly accurate. When compared to expert surgeon review, crowds of non-experts were able to predict the pass/fail rate of surgeons with 100% accuracy in retrospective studies [1, 19].

In addition to these human perception-based approaches, computational and event metrics have also predicted surgical technical skill with varying degrees of accuracy. Metrics such as Task Time (TT), spectral arc length (SAL), camera use per minute (CUP), normalized angular displacement (NAD), rate of orientation change (ROC), and mean velocity (MV), have all been found to associate with surgical skill [13, 27, 39, 40, 114].

In this study we sought to find the optimal VR warm-up module to properly prime surgeons for robotic surgery, and validate the module in the operating room (OR) by measuring objective performance metrics and obtaining video review of participants who were randomized to either participate in the module or not prior to surgery.

1.3 Methods

3.1 Aim 1

6.1.3.1.1 Subject Recruitment Surgeons and trainees were recruited after Insti- tutional Review Board (IRB) approval (UW IRB # 41730), from urology, gynecology and general surgery at the University of Washington Medical Center (UWMC), Madigan Army Medical Center (MAMC), and the Florida Hospital/Nicholson Simulation Center. Each of the subjects were asked to complete a standard demographics questionnaire to learn their experience and any factors impacting responses to warm-up simulations. The information obtained included:

- Unique Subject Code (e.g. UWMC03)
- Age
- Gender
- Years in Training
- Years since graduated training
- Surgical specialty
- Robotic surgery exp. (# cases as primary surgeon)
- Laparoscopic surgery exp. (# cases as primary surgeon)
- Handedness
- Musical experience
- Date of return since last deployment
- Duration of deployment
- Surgical role while deployed (if applicable)
- Time since last performed robotic surgery or procedure

Seattle Children's — Internal

6.1.3.1.2 dVSS Modules The modules used for surgeon priming needed to be complex and effective enough at stimulating the surgeon in a way that most closely corresponds to movements present in a surgical environment. The criteria used to select modules were: (1) prior validation in the literature, and/or 2) involvement of multiple skills simultaneously, and/or 3) containing content that closely simulates actual surgery. Following these criteria, we chose to use the Ring and Rail 2, and Match Board 3 modules from the EndoWrist Manipulation 2 category. Ring and Rail 2 has previously been found to effectively discriminate against surgical skills, and Match Board 3 has been found to be the most difficult of the dVSS modules among experienced surgeons [109, 115]. From the Needle Driving category, Suture Sponge 3 was chosen as it has been found to vary

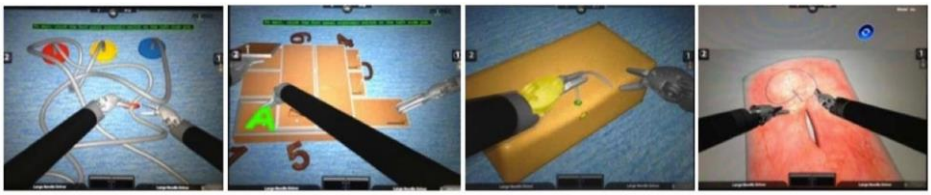


Figure 6.3: The four dVSS warm-up modules used for aim 1 of the RSR study including Ring and Rail 2, Match Board 3, Suture Sponge 3, and Running Suture (left to right).

6.1.3.1.3 greatly between experience, intermediate, and novice performers [115]. Running Suture was the final module chosen, and the criterion module, as it most closely represents the actions in a real surgical suturing procedure. Each of these four modules can be completed in approximately 2-3 minutes, making them short enough to be feasibly used as a component of a final curriculum (Fig. 6.3).

6.1.3.1.4 Proficiency Testing All recruited participants did not necessarily have the same levels of experience or skill. To train a benchmark of technical skill, each surgeon was required to complete the Intuitive Surgical da Vinci didactic web-based curriculum. Previous research has used this curriculum as a method to familiarize participants with the console and train surgeons to adequate proficiency levels [111]. Two expert robotic surgeons completed the four modules used, until they made no errors. These performances were used as the benchmark for expert skill.

6.1.3.1.5 The four dVSS modules were split into five groups with different combinations of modules, to test which module or module series was most efficient for priming surgeons (Fig. 6.4). The modules in each of these five groups are shown in Table 6.1. Participants were randomized to be in one of the 5 groups, and within each group, half were randomized to perform the tasks monthly and half performed bi-weekly. After 8 weeks of performing at one frequency, the participants were then swapped to the opposing

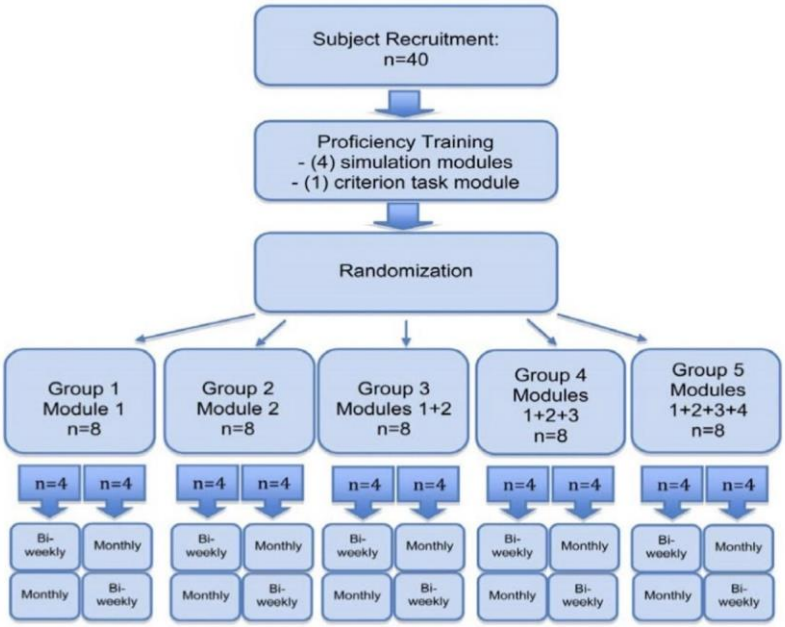


Figure 6.4: The randomization setup for the 5 warm-up groups in aim 1 of the RSR study.

6.1.3.1.6 frequency group, to allow for certain baselines in statistical models to control for intervals of inactivity. For groups with more than 1 module in a series, a 30 second period was inserted between each module performed, making the total simulation time approximately 3-12 minutes. After performing the module or module series, each participant performed the criterion module (Running Suture) in which they were assessed based on their performance by the two expert surgeons.

6.1.3.1.7

6.1.3.1.8 Video Review Two expert surgeons reviewed videos of the performances and rated the criterion suturing procedures. The GEARS assessment tool, being the most popular of the assessment tools for robotic surgery, was used as the evaluation method [89]. The GEARS evaluation consists of 6 Likert-scale domains which are relevant to robotic surgical technical skill. These domains are: depth perception, bimanual

Simulation Group	Group 1	Group 2	Group 3	Group 4	Group 5
Running Suture	X		X	X	X
Ring and Rail		X	X	X	X
Match Board				X	X
Suture Sponge					X

Table 6.1: The five groups of warm-up modules used for aim 1. Each group consisted of different combinations of warm-up modules shown to help elevate skill.

dexterity, force sensitivity, autonomy, efficiency, and robotic control. Each of these do- mains are given a score between 1 and 5, then added for a cumulative score between 6 and 30, with a higher score corresponding to a more skilled surgeon. Autonomy was removed from consideration for this study, as it can be used with a lone video review, leaving a total possible score of 5-25.

6.1.3.1.9 Statistical Models The primary goals of Aim 1 were to evaluate (1) the magnitude of skills decay between the two standardized assessment intervals, and (2) the curriculum which optimizes the participants GEARS scores and results in the smallest deviation from performer baseline performance.

The general model used was a linear mixed effects model which accounts for interval assessment period, timing of assessment, as well as the contrast between the different curriculum groups. This model included a random intercept to account for clustering of study participants, and a fixed effect to adjust for assessment time between different trials. To account for the average percentage of skill decay between groups, the contrast in GEARS scores between groups performing monthly and groups performing bi-weekly was accounted for in the model. The module which yielded the highest average combined GEARS score after accounting for all of these factors was used as the optimal warm-up curriculum for Aim 2.

3.2 Aim 2

6.1.3.2.1 Subject Recruitment Subjects from Aim 1 were allowed to participate in Aim 2 of the study, in addition to newly recruited participants which were approved by the IRB. Each newly enrolled subject in the second Aim completed the demographics questionnaire discussed in the section for Aim 1.

6.1.3.2.2 Randomization and Warm-Up Protocol Individual cases from each recruited surgeon were randomized to either receive the warm-up curriculum or not receive it, with each surgeon acting as their own control case. These assignments were stored centrally at the University of Washington Center for Biomedical Statistics (CBS) and distributed using a REDCap-based (http://project-redcap.org/) web delivery system. Each sites research coordinators ensured adherence to the warm-up protocol. Surgeons who were randomized to receive the warm-up were brought to a dVSS tower to complete the VR module selected in Aim 1. Participating surgeons completed be- tween 1-40 procedures as a part of this study. Each sites project coordinator worked with operating room staff to familiarize the operative team with the projects protocol to minimize disruptions to the daily operation routines.

6.1.3.2.3 OR Data Capture Endoscopic video and robotic instrument position was captured for each procedure using the Intuitive Surgical dvLogger system, which captures data directly from the robots API. Capturing this data allowed for using the criterion video with the GEARS assessment tool by expert surgeons and non-expert crowd workers, afterwards. This also allowed performance metrics to be computed using the obtained kinematic data from the criterion surgery.

6.1.3.2.4 Data Aggregation and Manipulation All video and tooltip data was sent in external hard drives to UMN for analysis. Files were then parsed into 'SI' or 'XI'labels based on which da Vinci platform was used. In several on-screen modules used to display which surgeon had control of which arms of the robot, the video in SI cases displayed personally identifiable information. These displayed modules were censored using OpenCV and Python to create black boxes over these areas.

6.1.3.2.5 Video Segmentation The videos in each surgical case were manually searched to find the timestamps at which the surgeon started operating. An effort was made for each video to obtain the first 15 minutes of surgery from the criterion surgeon, as this was the timeframe needed to see the largest effect from the warm-up module. Throughout several cases, the criterion surgeon was switching back and forth with a partner, requiring editing and splicing together several sections of video together to arrive at a 15 minute timeframe. An additional effort was made to remove all sections of video in which the endoscope leaves the patient's body, capturing possibly identifiable information. This video was then uploaded as "unlisted" to Youtube, to allow sharing videos with faculty surgeon reviewers.

6.1.3.2.6 Objective Metric Calculation The kinematic tooltip data was parsed using Python [87] and Intuitives internal API for decrypting the data. The dvLogger split surgeries chronologically into several directories in which a surgery could comprise over 100 separate directories, depending on the length of the surgery. A script was written to store each bin's kinematic data individually in a compressed format for later

Seattle Children's — Internal

analysis. An additional script was written to find the time intervals in the surgery when the arm's were moving to calculate tool path length. The information obtained from the raw data contained X, Y, Z coordinates in meters, along with a UNIX timestamp in milliseconds. The kinematic data and video were usually not well-synced so an effort was made to align the criterion portion of the video with the kinematic tooltip data by calculating the percentage of the video completed with the percentage of timestamps completed in the kinematic data, to get the closest possible match between the two modalities. These data were used to obtain summary metrics based on periods when the surgeon was performing, including rate of orientation change (ROC), spectral arc length (SAL), normalized angular displacement (NAD), mean velocity (MV), laterality (LR), and camera use per minute (CUP).

6.1.3.2.7 Video Review For faculty surgeon review, a web-based platform was created to distribute the processed endoscopic videos accompanied by the modified GEARS questionnaire to faculty surgeons and non-expert crowds. A subset of the total videos were used for faculty surgeon review, attempting to obtain reviews for as many videos as possible. Each GEARS score was recorded for each performance to analyze the intraclass correlation (ICC) among ratings for the same video. Statistical calculations were computed using R [92]. C-SATS Inc. was additionally used to compile quick and reliable reviews of technical skill for each video, obtained from crowds of non-experts, to offset the limited availability of faculty surgeons.

Results

3.3 Aim 1

Forty-one participants completed the study in Aim 1, although the original objectives were to only enroll forty, with demographics from these surgeons shown in Fig. 6.5. As can be seen in Fig. 6.6, the average TT for the different groups had a downward trend in which the surgeons became quicker as they proceeded through sessions. Clearly, the Ring and Rail 2" lone module, group 2, was the least effective at lowering TT while also, as seen in Fig. 6.7, resulted in the lowest GEARS scores. Groups 1 and 5 resulted in the most favorable TT and GEARS scores. Overall group 1 participants, the participants

	Florida	Madigan	Washington	Total
Target Enrollment	20	20	30	70
Total Randomized	**5 (25)**	**29 (145)**	**12 (40)**	**46 (66)**
Completed Participation	**3 (15)**	**27 (135)**	**11 (37)**	**41 (59)**
Consented*	7 (35)	38 (190)	29 (97)	74 (106)
Gender				
Female	1 (14)	10 (26)	9 (31)	20 (27)
Male	6 (86)	28 (74)	17 (59)	51 (69)
Other	0 (0)	0 (0)	0 (0)	0 (0)
Seniority				
Attending	7 (100)	12 (32)	3 (10)	22 (30)
Senior	0 (0)	9 (24)	9 (31)	18 (24)
Junior	0 (0)	17 (45)	14 (48)	31 (42)
Specialty				
Urology	3 (43)	12 (32)	12 (41)	27 (36)
Gynocology	1 (14)	9 (24)	2 (7)	12 (16)
General	2 (29)	8 (21)	10 (34)	20 (27)
Thoracic	0 (0)	0 (0)	0 (0)	0 (0)
Cardiothoracic	0 (0)	0 (0)	0 (0)	0 (0)
ENT	1 (14)	9 (24)	2 (7)	12 (16)
Handedness				
Left	1 (14)	1 (3)	1 (3)	3 (4)
Right	5 (71)	35 (92)	23 (79)	63 (85)
Ambidextrous	1 (14)	2 (5)	2 (7)	5 (7)
Musical Experience	2 (29)	22 (58)	19 (66)	43 (58)
Deployed	0 (0)	7 (18)	0 (0)	7 (9)

Figure 6.5: Enrollment demographics for participants in aim 1 of the RSR study.

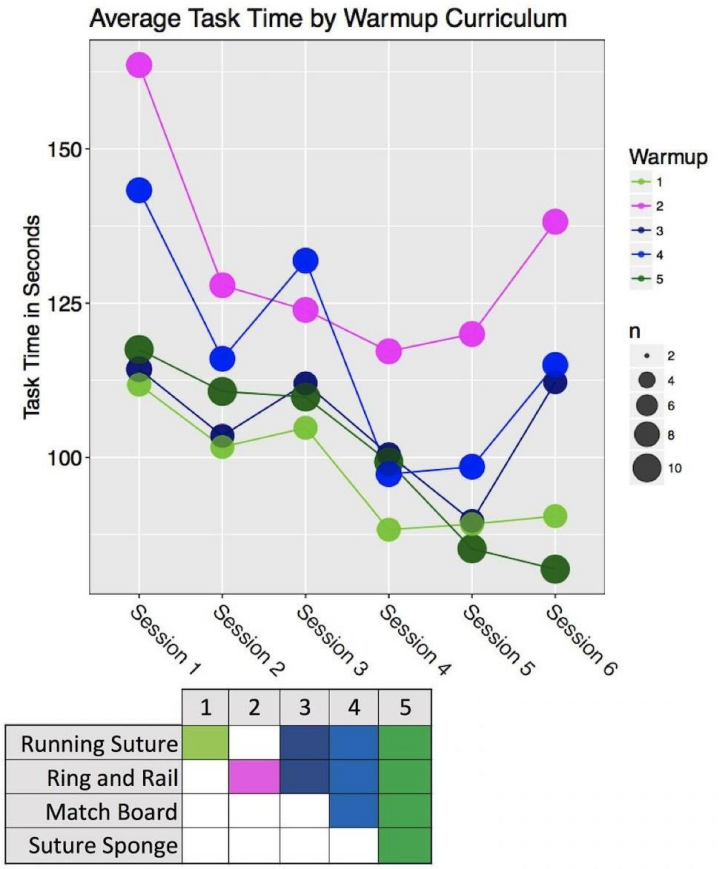


Figure 6.6: The task time (TT) for the five different warm-up module groups, through- out the six completed sessions.

only priming with the "Running Suture" module, had a higher GEARS score (2.1 points higher +/- 0.86, $p = 0.02$). Compared to group 2, group 1 was on average 31.3 seconds faster (+/- 14.02 seconds, $p = 0.03$). GEARS scores from groups 1 and 5 were not significantly different from each other ($p = 0.47$ and $p = 0.70$), although group 1 had the least variation in GEARS scores and TT of all groups. The criterion task, "Running Suture", thus was chosen as the warm-up curriculum to use before surgical procedures.

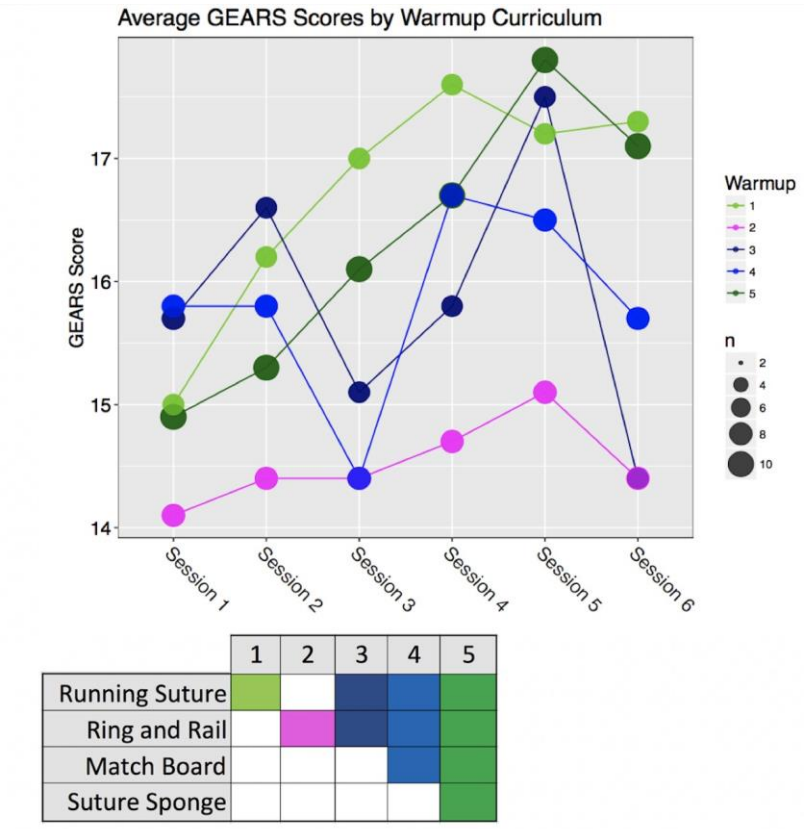


Figure 6.7: The GEARS scores for the five different warm-up module groups, throughout the six completed sessions.

	Madigan	Washington	Total
Sessions with Video/Kinematic	17	326	343
Completed Sessions	41	391	432
Consented*	12 (200)	22 (110)	34 (131)
Deployed Currently*	0 (0)	0 (0)	0 (0)
Lost to Follow-Up*	0 (0)	0 (0)	0 (0)
Gender			
Female	1 (8)	8 (36)	9 (26)
Male	10 (83)	11 (50)	21 (62)
Other	0 (0)	0 (0)	0 (0)
Specialty			
Urology	2 (17)	8 (36)	10 (29)
Gynocology	0 (0)	9 (41)	9 (26)
General	6 (50)	1 (5)	7 (21)
Thoracic	0 (0)	0 (0)	0 (0)
Cardiothoracic	1 (8)	0 (0)	1 (3)
Handedness			
Left	0 (0)	0 (0)	0 (0)
Right	11 (92)	19 (86)	30 (88)
Ambidextrous	0 (0)	0 (0)	0 (0)
Musical Experience	3 (25)	6 (27)	9 (26)
Deployed Previously	9 (75)	0 (0)	9 (26)

Figure 6.8: Enrollment demographics for participants in aim 2 of the RSR study.

3.4 Aim 2

Forty surgeons were enrolled in Aim 2, from UWMC and MAMC, with 34 surgeons having completed at least one surgery. Of the 34 surgeons, 30 answered the demographics questionnaire, with results shown in the Fig. 6.8. Video of 432 RAS were performed and recorded. Due to various technical malfunctions with the recording technology, the kinematic data for 89 cases were not recorded, leaving 343 videos having corresponding tooltip position kinematic data.

6.1.3.4.1 GEARS Review Outcome and the other five Ratings by seven faculty surgeons were obtained for 45 videos, with two raters rating all videos rating from 7

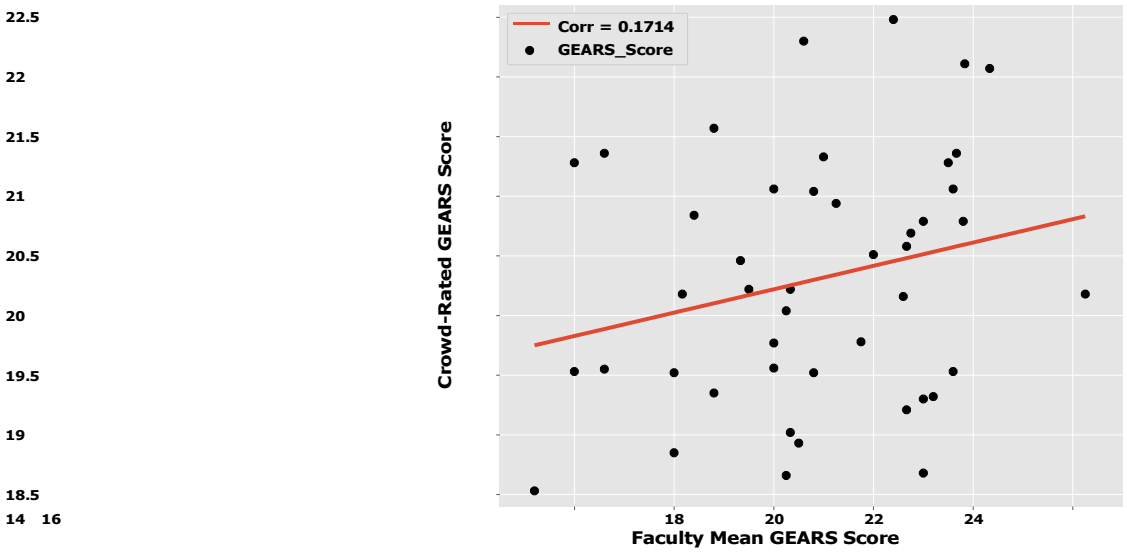


Figure 6.9: The correlation of the GEARS scores from faculty surgeons and C-SATS scores was weak, with a value of 0.1714, 95% CI from -0.1287 to 0.4426.

to 44 videos each. ICC for expert ratings was 0.3825 (95% CI from 0.2473 to 0.5366), indicating poor agreement. When the two raters with the greatest numbers of missing observations were removed, ICC was 0.4502 (95% CI from 0.3026 to 0.6043), still indicating poor agreement. ICC estimates and their confidence intervals were based on an individual ratings, consistency, 2-way random-effects model for incomplete datasets (using R package irrNA, Brueckl and Heuer, 2018). Crowdsourced GEARS scores sup- plied by C-SATS were moderately (but not significantly) correlated with the mean of the expert ratings for the same case (r(43) = 0.1714, 95% CI from -0.1287 to 0.4426, Fig. 6.9).

We assessed the effect of the intervention on GEARS score, our primary outcome, using both a linear mixed model with a random intercept for each surgeon (p = 0.9819) and a nonparametric modified Friedman test (p = 0.3441), with both finding no significant difference between cases receiving or not receiving the intervention. Robot type



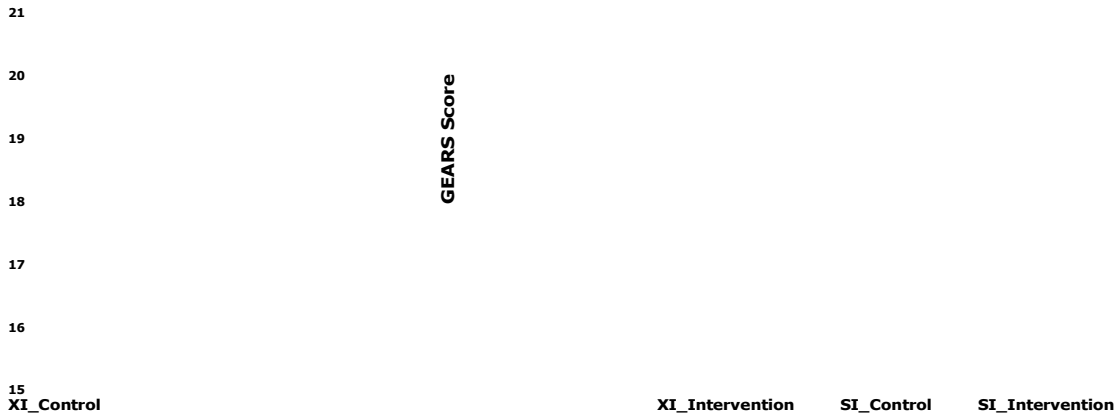


Figure 6.10: The effect of Si and Xi da Vinci platforms for surgeons who performed and did not perform a warm-up module before surgery. No significant difference exists between these four groups.

(XI vs SI) was also not significantly related to GEARS score ($p = 0.2365$), nor was any robot by treatment interaction detected ($p = 0.6694$, Fig. 6.10).

6.1.4.2.2 Objective Metric Outcomes We also examined the effect of the inter- vention on six kinematic and event outcomes (mean velocity (MV), rate of orientation change (ROC), normalized angular displacement (NAD), spectral arc length (SAL), camera use per minute (CUP), and laterality (LR)) using both linear mixed models and a nonparametric modified Friedman test, finding no significant differences between treatment arms. We used the Holm-Sidak procedure to maintain familywise alpha at .05.

Kinematic and event outcomes were not significantly associated with GEARS score as quantified by repeated measures correlation (using R package rmcorr [116]), (Table 6.2, and 6.3.

Metric	Estimated β	P-Value	P-Value Threshold
MV	0.0587	0.397	0.009
NAD	-0.063	0.426	0.010
SAL	0.059	0.655	0.013
CUP	0.027	0.766	0.017
LR	-0.009	0.790	0.025
ROC	0.021	0.888	0.050

Table 6.2: Results from linear mixed model tested with kinematic and event outcomes.

Metric	P-Value	P-Value Threshold
MV	0.3536	0.017
NAD	0.2909	0.010
SAL	0.3132	0.0127
CUP	0.6797	0.050
LR	0.4898	0.0253
ROC	0.1595	0.0085

Table 6.3: Results from Friedman test with kinematic and event outcomes.

Metric	Correlation	95% Upper CI	95% Lower CI	P-Value
MV	0.078	-0.031	0.187	0.161
NAD	-0.056	-0.165	0.054	0.315
SAL	-0.022	-0.132	0.088	0.692
CUP	0.079	-0.031	0.188	0.158
LR	-0.010	-0.119	0.100	0.864
ROC	-0.013	-0.123	0.097	0.811

Table 6.4: Correlation values for each of the metrics measured with GEARS scores.

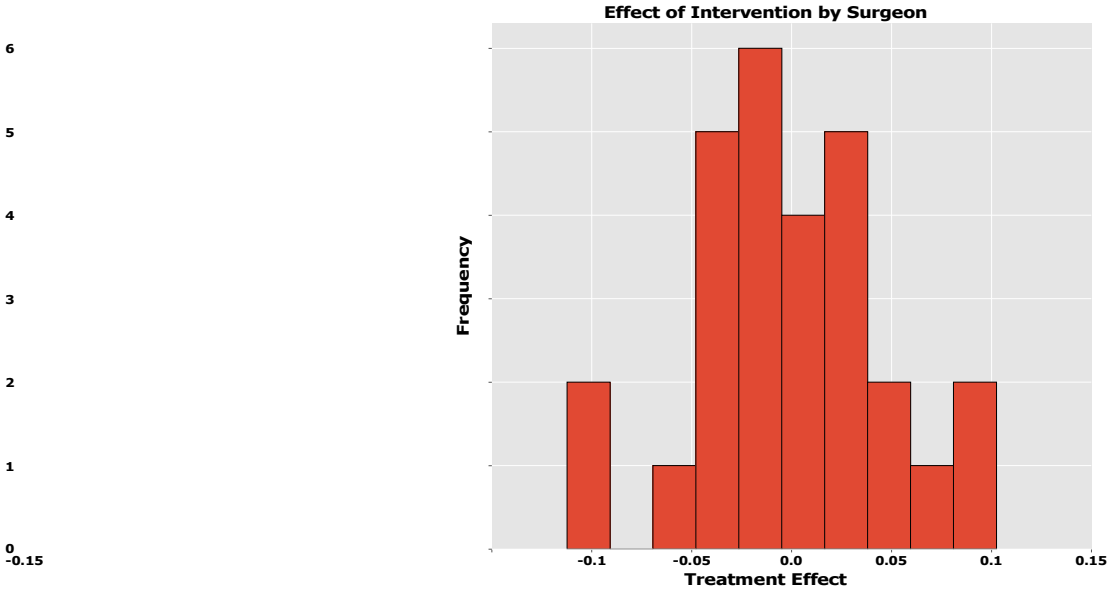


Figure 6.11: The effect of performing a warm-up module for each each surgeon in the RSR study. With a clear normal distribution, no significant effect was measured.

The median time between surgeries for surgeons in our study was 20 days. We did not find evidence for a difference in GEARS score ($p = 0.8331$) or treatment effect (p 0.5997) associated with greater time lapse between surgeries (Table 6.4.)

A linear mixed model including both a random intercept and a random slope for the effect of the treatment on each surgeon was used to quantify the range of effects the intervention had on different surgeons. Treatment effects for individual surgeons on GEARS score ranged from -.102 to .094 points, which we judged not to be clinically meaningful, thus precluding the possibility of identifying subgroups of surgeons who benefited from the intervention to any practically significant degree (Fig. 6.11).

1.4 Discussion

No statistically significant difference in perceived technical skill was observed be- tween robotic surgeons who perform a warm-up suturing module and those who do

not. Tests controlling for individual surgeons, Si vs. Xi da Vinci robot platforms, as well as time since last surgical performance, all found no evidence that performing the RSR curriculum prior to surgery elevated surgical readiness, as measured through both kinematic event metrics and faculty video review.

Efforts from this study led to the creation of the largest dataset (to the authors knowledge) of RAS with corresponding kinematic tooltip data to date. This could lead to breakthroughs in artificial intelligence models which map surgical technical skill to robotic surgery performance and video features. Large datasets are necessary for proper training of machine learning models which can both quantify technical skill and perform semantic segmentation. The rich, multimodal detail provided by this dataset, shown through both high definition video and the abundance of kinematic tooltip data should prove invaluable for future research.

1.5 Conclusions

Performing a VR warm-up module prior to robotic surgery does not significantly improve the technical skill or readiness of a robotic surgeon. The dataset created in this study will prove invaluable for future research including surgical skill classification, semantic segmentation, automated performance metric calculation, and surgical simulation research.

6.2 Temporal Variability in Surgical Technical Skill Evaluation [117]

2.1 Abstract

Purpose: Summary score metrics, either from crowds of non-experts, faculty surgeons, or automated performance metrics (APMs), have been trusted as the prevailing method of reporting surgeon technical skill. The aim of this paper is to learn whether there exist significant fluctuations in the technical skill assessments of a surgeon through- out long durations of surgical footage.

Methods: A set of 12 videos of common robotic surgery cases from human patient robotic surgeries were used to evaluate the perceived technical skill at each individual minute of the surgical videos, which were originally 12-15 minutes in length. A linear mixed effects model for each video was used to compare the ratings of each minute to the overall mean for the entire video in order to learn whether a change in scores over time is more significant than typical inter-rater variation.

Results: Modeling the change over time of the Global Evaluative Assessment of Robotic Skills (GEARS) scores significantly contributed to the prediction model. This gives evidence that significant changes in the technical skill over time are noticeable in the model.

Conclusion: The findings from this research raise questions about the optimal duration of footage needed to be evaluated to arrive at an accurate rating of surgical technical skill for longer procedures. This may imply non-negligible label noise for supervised machine learning approaches. In the future, it may be necessary to report a surgeon's skill variability in addition to their average score to have proper knowledge of a surgeon's overall skill level.

2.2 Introduction

Seattle Children's — Internal

Methods for assessing the technical skills of surgeons is paramount to ensuring to the public that

surgeons are safe and effective. For years summary scores or metrics have been used as the main method to report surgical skill. The most popular of these has been to use a likert-scale scoring metric, from either non-expert crowds or faculty surgeons. Past research has shown that crowds of non-experts concord with surgeon raters in evaluating technical skill [**?**]. Automated performance metrics have also been used, in which surgical events or streaming kinematic data have been used for computation of various metrics, suggesting superior objectivity [41]. As it is known that surgical skill is related to patient outcomes, and medical errors are the third leading cause of death in the United States, for which surgical errors contribute a large part, it remains important to accurately assess and report surgeon skill [4], [72]. Much progress has been made by using statistical and machine learning models in the past with the goal of classifying surgeons into skill levels of 'novice' and 'expert' [73], but it remains a difficult computational task. For the largest statistical power this usually leaves the most reliable approach being to obtain either surgeon or crowd evaluations from video [22].

One of the most popular robotic surgical skill assessment metrics is the Global Evaluative Assessment of Robotic Skills (GEARS), which is the most common objective assessment tool for robotic surgery [89]. The subdomains in this metric include: bimanual dexterity, efficiency, depth perception, force sensitivity and robotic control. Scores of 1-5 for each of these subdomains are totaled, for a cumulative score of 5-25, with higher scores belonging to more skilled surgeons. There is an additional domain in GEARS named Autonomy which is not used here. Autonomy is typically used when evaluating a surgeon's ability to work independently, which can't be accurately assessed through video alone.

Seattle Children's — Internal

At least two confounding factors arise: (1) limitation of human attention span of the raters which inhibits reliability of their ratings and (2) the fluctuations of "true technical skill", naturally exhibited by a surgeon through time, when videos of a longer duration are used. It is unclear whether ratings remain reliable as durations scale up. The former is akin to *measurement noise* and is an attribute of the method to measuring skill - be it human ratings or computed computationally. The latter is akin to *process noise* and is an attribute of the surgeon's activities and tool-tissue environment interactions. Studies of human attention span have found that people viewing video lectures on average experienced significant decreases in attention and even stopped viewing after only about 6 minutes. This study additionally found that including interactivity elements to video lectures led to an increase in watch time by at least 20% [118]. Given that evaluating surgeons requires an even higher degree of focus, using videos which are longer than 5 minutes may be detrimental to receiving accurate evaluations. Additionally, when making judgments, crowds can be susceptible to a contrast effect bias in which wide ranges in performance can lead to wildly different evaluations [119]. This could mean a performance of a novice who performed unusually well in the last segment of a video may receive unreliably high ratings. The natural fluctuations in technical skill remain largely unexplored, though hinted at in [91], [81].

We hypothesize that statistically significant degrees of temporal fluctuations of perceived skill exist between smaller segments of surgical performances, which differ from previously obtained summary scores of the longer duration videos.

2.3 Methods

3.1 Dataset

This study utilized 12 videos from the Robotic Surgery Readiness (RSR) study. The goal of that study was to determine whether pre-operative warm-up on virtual reality tasks had a measurable improvement in surgical technical skills among practicing surgeons (no novice trainees) after typical periods of surgical inactivity, in a controlled- randomized trial. This dataset contains 343 videos of surgical procedures consisting of live robotic surgeries, which were performed by forty attending surgeons and trainees in urology, gynecology, and general surgery at the University of Washington Medical Center and the Madigan Army Medical Center. These robotic surgeries were performed using da Vinci surgical robots, created by Intuitive Surgical (Sunnyvale, CA). Each video was manually edited to include roughly the first 15 minutes of surgical activity performed by the criterion surgeon. This beginning portion of the surgery contained similar actions among the different surgery specialties, such as suturing and cutting ac- tions. A GEARS score for each 12-15 minute performance was also previously obtained from crowd evaluation.

The range of scores in all RSR videos was fairly small, with most lying between 20-22 out of 25 (only 15% of the full range possible). This was the case as all surgeons assessed were practicing faculty and no trainees were involved in these videos. To obtain the largest possible range of skill from this dataset, 6 performances from the top quintile of scored performances and 6 from the bottom quintile of performances were used and given labels of 'expert' and 'proficient', respectively, keeping in mind that all surgeons in this dataset are highly-skilled, or they wouldn't be allowed to operate in live robotic surgery procedures [70]. Additionally, half of each group of videos were performances in which the surgeon participated in a simulated surgery warm up

Seattle Children's — Internal

procedure beforehand,

Grou	Sub-Group	Videos (N)	Prostatectomy	Hysterectomy	
Expert	Warmup	3	1	1	1
	Control	3	1	0	2
Proficient	Warmup	3	2	0	1
	Control	3	1	1	1

Table 6.5: Summary information for 12 videos used from the RSR study.

whereas the other half started surgery without warming up. This stratified formatting is illustrated in Table 6.5.

3.2 Crowd Evaluation

Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each non-expert crowd worker was paid an average of $0.50 to watch and evaluate each of a series of one minute videos. A web domain was created for which Turkers would be redirected, where they submitted a consent form and were asked GEARS questions about videos. Each 12-15 minute video performance was segmented into smaller videos having a duration of one minute, using FFMPEG, an open source video editing tool [85]. The order of these videos were randomized so that crowds were viewing a non- chronological ordering. They were asked to provide a GEARS score for each minute long video and proceeded to evaluate all remaining videos for the same performance.

A few extra elements were added to the user interface in an effort to increase interactivity and obtain better data quality. Once a crowd worker gave consent to participating in the study, they were shown a video with two performances, side by side, of a novice and expert performer of a laparoscopic surgical training task. The crowds were asked to evaluate which surgeon was better, and to what degree. After evaluating this task, the series of robotic surgery procedures began. After each GEARS likert-scale question

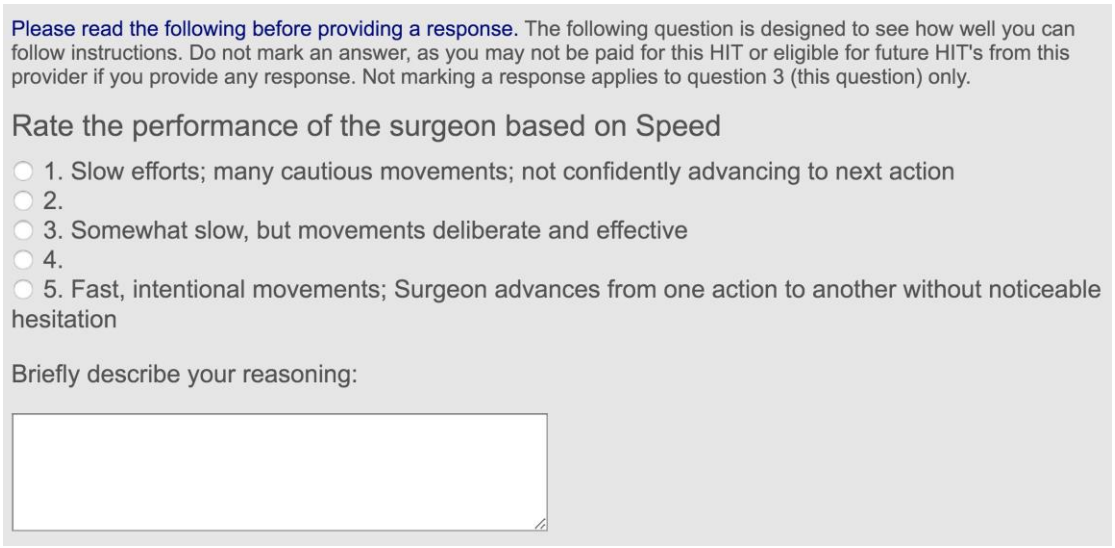


Figure 6.12: The attention question inserted to the GEARS questionnaire, for quality assurance purposes. Note that skimming the instructions likely results in incorrect answers.

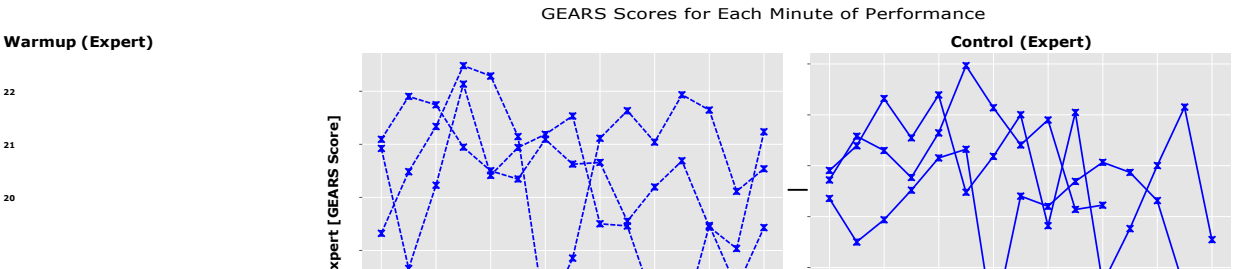
prompt, a text box was inserted, asking the raters to provide reasoning for their decision in each subdomain.

Finally, an additional question, which is not part of the standard GEARS questionnaire was inserted, asking for a rating of the 'speed' of the surgeon, shown in Fig.6.12. Text in the prompt stated that the worker should not answer this question. This 'attention question' served as a mechanism to query which turkers were focusing on the questionnaire vs. those quickly finishing to be paid. All ratings which included answers to this 'attention question' were removed from analysis.

3.3 Statistical Analysis

A linear mixed effects model for each video was used to analyze the outcome of GEARS scores assigned by each rater at each timepoint. A random intercept was estimated for each individual rater to model rater-by-rater variability in scores. Fixed effects were also estimated for each minute of video in order to quantify differences in surgical skill over time. With each model, we tested the hypothesis that the mean GEARS score was the same at each timepoint to determine if there were significant changes in skill over time for each surgeon. All statistical work was accomplished using R 3.6.2 [92], with data manipulation and visualization computed in Python 3.6 [87].

2.4 Results



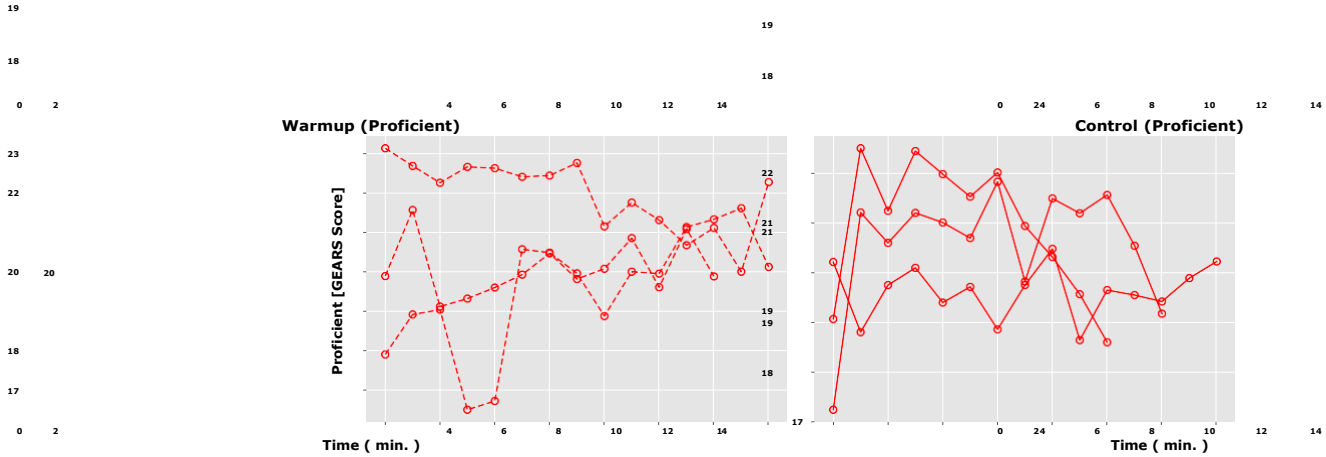


Figure 6.13: All mean crowd evaluations from each proficient (red) and expert (blue) surgeon at each minute of the performance. Surgeons previously randomized to the control (solid lines) group normally started surgery without any intervention or pre- operative warm-up used. Warm-up (dashed lines) group surgeons reviewed a virtual reality warm-up module prior to surgery. (The warm-up hypothesis from the original randomized study is not being tested or evaluated in this research, only temporal variation in ratings is.

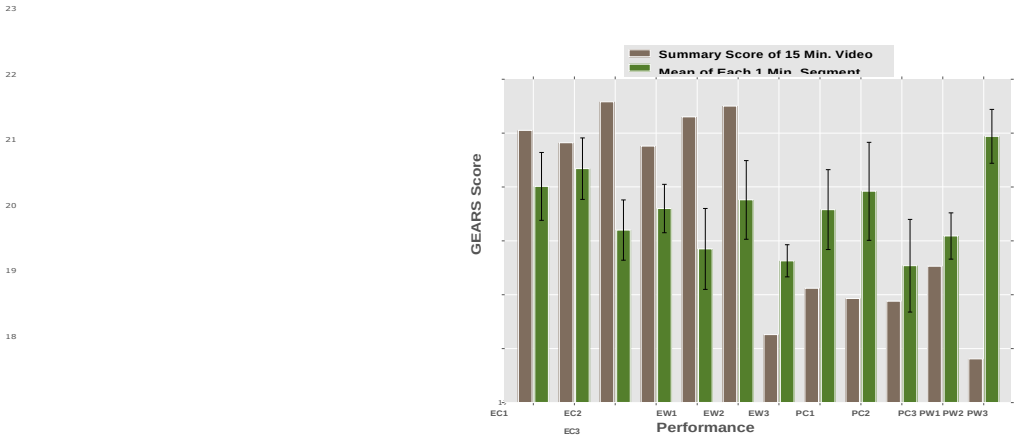


Figure 6.14: Comparison of previously-obtained scores for 15 minute video (95% CI) and average score for evaluation of 1 minute segment. (PW = Proficient/Warmup; PC = Proficient/Control; EW = Expert/Warmup; EC = Expert/Control)

The mean of the GEARS score for each expert and proficient surgeon are shown in Fig. 6.13. No identifiable relationship or trend in any of the 4 subplots is immediately apparent. However, significant differences in the average score over time were detected for ten of the twelve videos (p = 0.001891 to $2.531e-14$) using linear mixed effects models. For two videos, the effect of time was not significant (proficient control video 1, p = 0.61794, and proficient warm-up video 2, p = 0.2618).

A bar plot comparing the previously obtained summary score for each surgeon, grouped with the average of each 1 minute segment, is displayed in Fig. 6.14. Several of the previously obtained summary scores are outside the confidence interval band from the average of the one minute scores. Overall we saw each full duration video's 1-minute GEARS average scores vary by 1.02 ± 0.28 points.

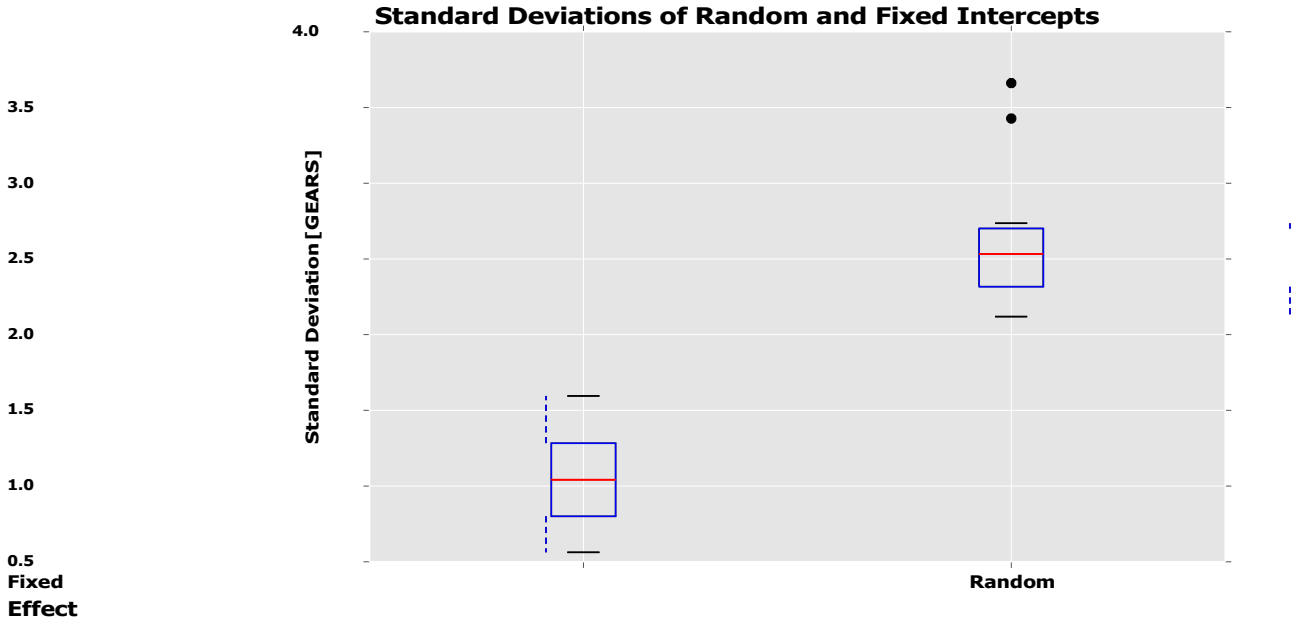


Figure 6.15: Standard deviation from the LME model for random effects from each individual crowd rater, compared with the fixed effect of time. The rater effect is clearly larger, although time is still significant in predicting outcomes.

6.2.5 Conclusion

The results from the video segmentation study support our initial hypothesis that perceived surgeon skill may fluctuate throughout surgical procedures. If we consider the limitations of human attention span and time-varying fluctuations in skill for a 15-minute video as contributing factors, it appears that human attention span limitations have a larger effect than natural variation of "true" technical skill in time (i.e. measurement noise may be 2-3x larger than process noise). We found no compelling evidence that the underlying human skill (process) is constant for a 15-minute duration in typical surgery. Perhaps crowds may be biased into giving scores that more closely reflect events they remember as being particularly good or bad. This research may give credit to alternative surgical skill reporting methods, as opposed to giving static scores

and labels. One alternative to this could be providing the mean of shorter segments of videos, as well as the amount of deviation in score throughout the performance, as a confidence interval. This would convey both the typical skill of a surgeon in addition to how consistent their skill is through time.

This study includes a few limitations. There was no semantic segmentation of these tasks, only temporal "chunking" into one minute segments. Although videos of real surgeries were used, the data used was fairly sparse, with only 12 videos of 15 minute surgeries. It is also not clear if the one minute duration used for segmentation is optimal. Shorter or longer segments of surgery may be more beneficial for accurate skill evaluation. In this work, the one minute ratings were obtained 3 months after the full duration videos were rated. This may raise concerns about reliability (and comparability of data) of crowd ratings from studies separated by large time intervals.

Using these alternative reporting methods could aid supervised machine learning models in the future by reducing potential label noise. Static scores may lead to noisily- labeled data sets, producing poorly trained classification models. Further study should consider optimal duration for human raters and a more rigorous analysis of the natural fluctuations of human technical skill.

Chapter 7

Conclusion and Discussion

This thesis presents work pertaining to multiple facets of how computational techniques in evaluating surgeon skill are affected by speed and task time. This work includes two research studies with conclusive evidence of biases to speed in skill evaluation efforts, as well as analysis surrounding how surgeon's hand movements correlate to expert skill and specific techniques which may be incorporated to classify surgical skill. In addition, this work includes findings from a new robotic surgery dataset in which it was found that performing a running suture virtual reality simulation prior to a robotic surgery did not lead to an elevated level of robotic technical skill. A secondary finding from this dataset shows the existence of temporal fluctuations in skill throughout longer surgical procedures. These findings help uncover different techniques of controlling for task time in surgical skill evaluation as well as uncovering different ways current evaluation techniques can be biased to speed.

Seattle Children's — Internal

7.1 Contributions

- An exploration of past surgical skill evaluation contributions by the research com- munity, including techniques which have been promising and areas where speed can effect calculations.
- A machine learning algorithm capable of classifying surgical performances with state-of-the-art accuracy in a manner which controls for time of task completion using tooltip kinematic position data.
- An in-depth examination of speed perception defining the manner in which non- expert human raters are biased to perceive surgeons as performing better when they appear to be moving faster, in a manner which doesn't extend to novice surgeons.
- A deeper examination of a sub-set of highly skilled robotic surgeons, examining how this group of surgeons is perceived as even more proficient when they appear to be making quicker movements.
- A statistical analysis of velocity profile models for surgeon hand movements, find- ing that expert surgeons are *less* likely to follow these models in their movements.
- A dataset of over 340 de-identified and labeled robotic surgery cases which consist of high quality video as well as tooltip kinematic data recorded from da Vinci surgical robots.
- A study finding that performing a virtual reality warm-up module prior to robotic surgery did not significantly improve the skill of surgeons, as measured by motion metrics, non-expert crowds, and faculty surgeons.
- An analysis exploring the temporal fluctuations in surgical skill, questioning the standard reporting metrics in the field of surgical skill evaluation.

Seattle Children's — Internal

.2 Limitations

There are a number of limitations to be noted from this work.

7.2.0.0.1 **Speed Perception Studies** The initial study which used dry lab laparo- scopic task videos used only ten videos, with five surgeons in each of the two skill groups. It is possible that if more statistical power were obtained by using a dataset with a larger volume of novices, the negligible increase in scores at quicker playback speeds for novice surgeons would no longer appear.

7.2.0.0.2 **Long Short-Term Memory Study** First and foremost, these tasks were simulated procedures, and the proposed algorithms and techniques may perform differently on real surgeries. The proposed method also only analyzes tool motion data, which may not contain sufficient data required for complete skill classification. The results from the crowd reassessment could be in part due to differences in the user interface of the skill evaluation web pages used for the two different occurrences of testing.

7.2.0.0.3 **Robotic Surgery Readiness Study** The RSR project consisted of only robotic, minimally invasive surgeries. No laparoscopic or open surgeries were used. It is possible that skills needed for robotic surgery are fundamentally different than other surgical fields, causing no warm-up effect to be noticeable in this study. Only five warm-up modules were assessed as possible warm-ups to use for this study.

7.2.0.0.4 **Rating Variability Study** There was no semantic segmentation of these tasks, only temporal "chunking" into one minute-segments. Although videos of real surgeries were used, the dataset was fairly sparse, with only 12 videos of 15 minute surgeries. It is also not clear if the one minute duration used for segmentation is the optimal duration to use in obtaining crowd ratings. Shorter or longer segments of surgery may be more beneficial for accurate skill evaluation. In this work, the one minute ratings were obtained 3 months after the full duration videos were rated. This may raise concerns about reliability (and comparability) of crowd ratings from studies separated by large time intervals.

7.3 Future of Surgical Skill Research

Several institutions are working towards developing new techniques to evaluate the skill of a surgeon. Extra care should be taken in the future to prevent creating techniques with biases towards faster surgeons, and instead use that data as one aspect of a broader list of attributes to evaluate. Making datasets open to the public for collaboration in skill evaluation research is the most beneficial way to innovate and build a safer surgical system, in which the annual number of medical errors is severely decreased. In the future, the creators of the

Seattle Children's — Internal

datasets mentioned in this thesis plan to make them publicly available for other research institutions to use. Furthermore, large professional societies of surgeons should be educated to use objective methods of evaluating skill, as opposed to subjective opinions. The collaboration of clinicians and scientists is a needed step forward to create a better future for health care.

Seattle Children's — Internal

References

[1] D. Holst, T. Kowalewki, L. White, T. Brand, J. Harper, M. Sorenson, S. Kirsch, and T. Lendvay. Crowd-sourced assessment of technical skills: An adjunct to urology resident surgical simulation training. *Journal of Endourology*, 29(5):604– 609, 2015.

[2] A. Jacobs, J. Pinto, and M. Shiffrar. Experience, context, and the visual percep- tion of human movement. *Journal of Experimental Psychology: Human Perception and Performance*, 30:822–835, 2004.

[3] T. Flash and N. Hogan. The coordination of arm movements an experimentally confirmed mathematical model. *The Journal of Neuroscience*, 5(7):1688–1703, 1985.

[4] M. Makary and M. Daniel. Medical error - the third leading cause of death in the us. *The BMJ*, 2016.

[5] L. Kohn, J. Corrigan, and M. Donaldson. To err is human. *National Academy Press*, 2000.

I. J. Rosen, M. Salazzo, B. Hannaford, and M. Sinanan. *Medicine Meets Virtual Reality*, 2001.

[6] C.E. Reiley and G.D. Hager. Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. *Medical Image Computing and Computer- Assisted Intervention*, pages 435–442, 2009.

[7] G. Megali, S. Sinigaglia, O. Tonet, and P. Dario. Modelling and evaluation of sur- gical performance using hidden markov models. *IEEE Transactions on Biomedical Engineering*, 53(10), 2006.

[8] M. Panzner and P. Cimiano. Comparing hidden markov models and long short term memory neural networks for learning action representations. *Semantic Com- puting Group*.

[9] H. I. Krebs, M.L. Aisen, B.T. Volpe, and N. Hogan. Quantization of continuous arm movements in humans with brain injury. *Proceedings of the National Academy of Sciences*, 96(2):4645–4649, April 1999.

[10] B. Rohrer and N. Hogan. Avoiding spurious submovement decompositions ii: a scattershot algorithm. *Biological Cybernetics*, 94:409–414, March 2006.

[11] C. Atkeson and J. Hollerbach. Kinematic features of unrestrained vertical arm movements. *Journal of Neuroscience*, 5(9):2318–2330, 1985.

[12] T.M. Kowalewski, L.W. White, T.S. Lendvay, I.S. Jiang, R.S. Sweet, A. Wright, B. Hannaford, and M.N. Sinanan. Beyond task time: automated measurements augments fundamentals of laparoscopic skills methodology. *Journal of Surgical Research*, 192(2):329–338, 2014.

[13] et. al. K. Ghani, B. Comstock. Technical skill assessment of surgeons performing robot-assisted radical prostatectomy: Relationship between crowdsourced review and patient outcomes. In *American Urological Association*, Boston, MA., 2017.

[14] L. Hammon and H. Hippner. Crowdsourcing. *Business and Information Systems Engineering*, 4(27):163–166, 2012.

[15] James Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[16] F. Galton. Vox populi. *Nature*, 75:450–451, 1907.

[17] A. Carvalho, S. Dimitrov, and K. Larson. How many crowdsourced workers should a requester hire. *Annals of Math and Artificial Intelligence*, 78(1):45–72, 2016.

[18] T.M. Kowalewski, B. Comstock, R. Sweet, C. Schaffhausen, A. Menhadji, T. Averch, G. Box, T. Brand, M. Ferrandino, J. Kaouk, B. Knudsen, J. Landman,

B. Lee, B.F. Schwartz, E. McDougall, and T.S. Lendvay. Crowd-sourced assess- ment of technical skills for validation of basic laparoscopic urologic skills tasks. *The Journal of Urology*, 195(6):1859–1865, 2016.

[19] Inc. CSATS. Validation, 2019.

[20] L. White, T. Kowalewski, R. Dockter, B. Comstock, B. Hannaford, and T. Lend- vay. Crowd-sourced assessment of technical skill (c-sats): A valid method for discriminating basic robotic surgery skills. *Journal of Endourology*, 29, 2015.

[21] C. Chen, L. White, T. Kowalewski, R. Aggarwat, C. Linnot, B. Comstock, K. Kuk- senok, C. Aragon, D. Holst, and T. Lendvay. Crowd-sourced assessment of tech- nical skills: a novel method to evaluate surgical performance. *Journal of Surgical Research*, 187(1):65–71, 2014.

[22] N. Aghdasi, R. Bly, L. White, B. Hannaford, K. Moe, and T. Lendvay. Crowd- sourced assessment of surgical skills in cricothyrotomy procedure. *Journal of Surgical Research*, 196(2):302–306, 2015.

[23] M.C. Vassiliou, L.S. Feldman, C.G. Andrew, S. Bergman, K. Leffondre, D. Stan- bridge, and G.M. Fried. A global assessment tool for evaluation of intraoperative laparoscopic skills. *The American Journal of Surgery*, 190(1):107–113, 2005.

[24] R. Plamondon, A. M. Alimi, P. Yergeau, and F. Leclerc. Modelling velocity profiles of rapid movements: a comparitive study. *Biological Cybernetics*, 69:119– 128, January 1993.

[25] N.E. Berthier and R. Keen. Development of reaching in infancy. *Experimental Brain Research*, 169:507–518, December 2005.

[26] S. Balasbramanian, A. Melendez-Calderon, and E. Burdet. A robust and sensitive metric for quantifying movement smoothness. *IEEE Transactions of Biomedical Engineering*, 59(8):2126–2136, August 2012.

[27] S. Balasubramanian, A. Melendez-Calderon, A. Roby-Brami, and E. Burdet. On the analysis of movement smoothness. *Journal of NeuroEngineering and Rehabil- itation*, 112, December 2015.

[28] A. Georgopoulos, J. Kalaska, and J. Massey. Spatial trajectories and reaction times of aimed movements: Effects of practice, uncertainty, and change in target location. *Journal of Neurophysiology*, 46(4):725–743, 1981.

[29] P. Morasso. Spatial control of arm movements. *Experimental Brain Research*, 42:223–227, 1981.

[30] S.W. Keele. Movement control in skilled motor performance. *Psychological Bul- letin*, 70(6):387–403, 1968.

[31] W.S. McCulloch and W.H. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematic Biophysics*, 5:115–133, 1943.

[32] B. Widrow. Adaptive sampled-data systems — a statistical theory of adaptation. *IRE WESCON Convention Record*, 4:74–85, 1959.

[33] Simple model of a small artificiaal neural network, 2017.

[34] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[35] A. Graves, N. Jaitly, and A. Mohamed. Hybrid speech recognition with deep bidirectional lstm. *IEEE Workshop on Automatic Speed Recognition and Under- standing*, 2013.

[36] H. Soltau, H. Liao, and H. Sak. Neural speech recognizer: Acoustin-to-word lstm model for large vocabulary speech recognition. *arXiv:1610.09975*, 2016.

[37] Siri Team. Deep learning for siri's voice: On-device deep mixture density networks for hybrid unit selection synthesis, 2017.

[38] J. Chen, P.J. Oh, N. Cheng, A. Shah, J. Montez, A. Jarc, L. Guo, I.S. Gill, and A.J. Hung. Use of automated performance metrics to measure surgeon perfor- mance during robotic vesicourethral anastomosis and methodical development of a training tutorial. *Journal of Urology*, 200(4):895–902, 2018.

[39] J.H. Nguyen, J. Chen, S.P. Mashall, S. Ghodoussipour, A. Chen, I.S. Gill, and A.J. Hung. Using objective robotic automated performance metrics and task-evoked pupillary response to distinguish surgeon expertise. *World Journal of Urology*, 2019.

[40] A. Hung, J. Chen, T. Nilanon, A. Jarc, M. Titus, P.J. Oh, I.S. Gill, and Y. Liu. Utilizing machine learning and automated performance metrics to evaluate robot- assisted radical prostatectomy performance and predict outcomes. *Journal of Endourology*, 32(5), 2018.

[41] A. Zia, C. Zhang, X. Xiong, and A.M. Jarc. Temporal clustering of surgical activities in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 12(7):1171–1178, 2017.

[42] A. Zia, Y. Sharma, V. Bettadapura, E.L. Sarin, M.A. Clements, and I. Essa. Au- tomated assessment of surgical skills using frequency

Seattle Children's — Internal

analysis. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. 18th Interna- tional Conference*, volume 9349, pages 430–438, Munich, Germany, September 2015.

[43] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selec- tion. *Pattern Recognition*, 3(15):1119–1125, November 1994.

[44] H. Wang, M. Muneeb Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[45] I. Laptev. On space-time interest points. *Internation Journal of Computer Vision*, 64(2):107–123, 2005.

[46] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, June 23-28 2008. IEEE Xplore, https://ieee.org. Date of Access: April 4, 2018.

[47] Fundamental of features and corners.

[48] Stanford University. Introduction to computer vision - tutorial 2: Image matching.

[49] Y. Li, R. Xia, Q. Huang, W. Xie, and X. Li. Survey of spatio-temporal interest point detection algorithms in video. *IEEE Access*, 5:10323–10331, 6 2017.

[50] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. *British Machine Vision Confer- ence*, September 2009.

[51] Y. Sharma, V. Bettadapura, N. Hammerla, S. Mellor, R. Mcnaney, P. Olivier, E. Deshmukh, A. Mccaskie, and I. Essa. Video based assessment of osats using sequential motion textures. *Georgia Institute of Technology Library*, 2014.

[52] A. Zia and I. Essa. Automated surgical skill assessment in rmis training. *In- ternational Journal of Computer Assisted Radiology and Surgery*, 13(5):731–739, March 2018.

[53] A. Zia, C. Zhang, X. Xiong, and A.M. Jarc. Temporal clustering of surgical activities in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 12(7):1171–1178, July 2017.

[54] Y. Sharma, V. Bettadapura, T. Plotz, N. Hammerla, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, and I. Essa. Video based assessment of osats using sequential motion textures. In *Fifth Workshop on Modeling and Monitoring of Computer Assisted Intervations – M2CAI*, 2014.

[55] A. Zia, Y. Sharma, V. Bettadapura, E.L. Sarin, and I. Essa. Video and accelerometer-based motion analysis for automated surgical skills assessment. 2017.

[56] M.J. Fard, S. Ameri, E.R. Darin, R.B. Chinnam, A.K. Pandya, and M.D. Klein. Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *International Journal of Medical Robotics*, 14(1), 2018.

[57] I. Funke, S.T. Mees, J. Weitz, and S. Speidel. Video-based surgical skill assessment using 3d convolutionaal neural networks. 2019.

[58] Z. Wang and A.M. Fey. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. 2019.

[59] L.T. Kozlowski and J.E. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception and Psychophysics*, 21:575–580, 1977.

[60] P. Veto, W. Einhäuser, and N.F. Troje. Biological motion distorts size perception. *Scientific Reports*, 7(42576):1–6, 2017.

[61] C. D Barclay, J.E. Cutting, and L.T. Kozlowski. Temporal and spatial factors in gait perception that influence gender recognition. *Perception and Psychophysics*, 23(2):145–152, 1978.

[62] R. Blake and M. Shiffrar. Perception of human motion. *Reviews in Advance*, 15(8), 2007.

[63] J.A. Beintema, K. Georg, and M. Lappe. Perception of biological motion from limited-lifetime stimuli. *Perception and Psychophysics*, 68:613–624, 2006.

[64] M.A. Giese and M. Lappe. Measurement of generalization fields for the recognition of biological motion. *Vision Research*, 42:1847–1858, 2002.

[65] D. Jokisch and N.F. Troje. Biological motion as a cue for the perception of size. *The Journal of Vision*, 3(4):252–264, 2003.

[66] J. Bruin. Introduction to linear mixed models.

[67] Newsom. Distinguishing between random and fixed: Variables, effects, and coef- ficients, 2019.

[68] J.D. Kelly, A. Petersen, T.S. Lendvay, and T.M. Kowalewski. Bidirectional long short-term memory for surgical skill classification of temporally segmented tasks. *International Journal of Computer Assisted Radiology and Surgery*, SUBMIT- TED.

[69] R. Dockter, T.S. Lendvay, R.M. Sweet, and T.M. Kowalewski. The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. *International Journal of Computer-Assisted Radiology and Surgery*, 12:1151–1159, 2017.

[70] H.C. Lin, I. Shafran, T.E. Murphy, A.M. Okamura, D.D. Yuh, and G.D. Hager. Automatic detection and segmentation of robot-assisted surgical motions. *Medical Image Computing and Computer-Assisted Intervention*, 3749, 2005.

[71] J.D. Birkmeyer, J.F. Finks, A. O'Reilly, M. Oerline, A.M. Carlin, A.R. Nunn, J. Dimick, M. Banerjee, and N.J. Birkmeyer. Surgical skill and complication rates after bariatric surgery. *New England Journal of Medicine*, 369(15):1434–1442, 2013.

[72] J.D. Kelly, A. Petersen, T.S. Lendvay, and T.M. Kowalewski. The effect of video playback speed on surgeon technical skill perception. *International Journal of Computer Assisted Radiology and Surgery*, 2020.

[73] M. Schuster and K.P. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(5), 1997.

[74] A.M. Derossis, G.M. Fried, M. Abrahamowicz, H.H. Sigman, J.S. Barkun, and J.L. Meakins. Development of a model for training and evaluation of laparoscopic skills. *American Journal of Surgery*, 175:482, 1998.

[75] G.M. Fried. Fls assessment of competency using simulated laparsocopic tasks. *Journal of Gasteroenterology Surgery*, 12:210, 2008.

[76] J.H. Peters, G.M. Fried, L.L. Swanstrom, N.J. Soper, L.F. Silin, B. Schirmer, and K. Hoffman. Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. *Surgery*, 135:21.

[77] R.M. Seete, R. Beach, F. Sainfort, P. Gupta, T. Reihsen, L.H. Poniatowski, and E.M. McDougall. Introduction and validation of the american urological associa- tion basic laparoscopic urology surgery skill curriculum. *Journal of Endourology*, 26:190.

[78] T.M. Kowalewski, R. Seete, T.S. Lendvay, A. Menhadji, T. Averch, G. Box, T. Brand, M. Ferrandino, J. Kaouk, B. Knudsen, J. Landman, B. Lee, B.F. Schwartz, and E. McDougall. Validation of the aua blus tasks. *Journal of Urology*, 195:998, 2016.

[79] A. French, K. Seidel, T.S. Lendvay, and T.M. Kowalewski. Role of contextual information in skill evaluation of minimally invasive surgical training procedures. *Hamlyn Symposium on Medical Robotics*, 2018.

[80] J.D. Kelly, Nicholas Heller, A. Petersen, T.S. Lendvay, and T.M. Kowalewski. The effect of video playback speed on robotic surgery skill evaluation. *International Journal of Computer Assisted Radiology and Surgery*, SUBMITTED.

[81] A.A. Gumbs, N.J. Hogle, and D.L. Fowler. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *Journal of the American College of Surgeons*, 204(2):308–313, 2007.

[82] A. Bajcsy, D.P. Losey amd M.K. O'Malley, and A.D. Dragan. Learning from phys- ical human corrections, one feature at a time. *In Proceedings of 2018 ACM/IEEE International Conference on Human-Robot Interaction*.

[83] T.M. Kowalewski. Real-time quantitative assessment of surgical skill. *PhD Thesis, University of Washington*, 2012.

[84] FFMPEG Developers, version 4.1.3 [software] edition, 2016.

[85] P. Holoborodko.Smooth noise robust differentiators. `http://www.holoborodko.com/pavel/numerical-methods/numerical-derivative/ smooth-low-noise-differentiators/`, year = 2008.

[86] Python Software Foundation. *Python Language Reference*. `http://www.python. org`, version 3.6 [software] edition.

[87] *MATLAB R2019a [Software]*. The Mathworks Inc., Natick, Massachusetts.

[88] F. C. Goh, D.W. Goldfarb, J.C. Sander, B.J. Miles, and B.J. Dunkin. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. *The Journal of Urology*, 187(1):247–252, 2012.

[89] R.M. Satava, A.G. Gallagher, and C.A. Pellegrini. Surgical competence and sur- gical proficiency: definitions, taxonomy, and metrics. *Journal of the American College of Surgeons*, 196(6):933–7, 2003.

[90] A. French, T.S. Lendvay, R.M. Sweet, and T.M. Kowalewski. Predicting surgical skill from the first n seconds of a task: Value over task time using the isogony principle. *The International Journal of Computer Assisted Radiology and Surgery*, 12(7):1161–1170.

[91] R Core Team. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org/, Vienna, Austria, 2018 edition.

[92] J.D. Kelly, A. French, M. Nash, L. Meryman, N. Heller, N. Organ, E. George, R. Smith, M. Sorensen, T. Brand, B. Comstock, T.M. Kowalewski, and T.S. Lendvay. Robotic surgery readiness: A randomized controlled study of virtual reality warm-up prior to robot-assisted surgery. *Journal of the American College of Surgeons*, In Preparation.

[93] C. Zhan and M. Miller. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. *JAMA*.

[94] S. Murhpy, J. Xu, and K. Kochanek. Preliminary data for 2010. *National Vital Statistics Reports*, 60(4), 2012.

[95] J. Shreve. The economic measurement of medical errors. *Society of Actuaries*, 2010.

[96] G. Sroka, L. Feldman, and M. Vassiliou. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the oper- ating room - a randomized controlled trial. *American Journal of Surgery*, 199:115, 2010.

[97] D. Stefanidis, J.R. Korndorffer Jr., R. Sierra, C. Touchard, J.B. Dunne, and D.J. Scott. Skill retention following proficiency-based laparoscopic simulator training. *Surgery*, 138:165–170, 2005.

[98] R.S. Perez, A. Skinner, P. Weyhrauch, J. Nehaus, C. Lathan, S.D. Schwaitzberg, and C.G. Cao. Prevention of surgical skills decay. *Military Medicine*, 178:588–592, 2013.

[99] K.H. Sheetz, J. Claflin, and J.B. Dimick. Trends in the adoption of robotic surgery for common surgical procedures. *JAMA Network Open*, 3(1), 2020.

[100] A.B. Haynes, T.G. Weiser, W.R. Berry, S.R. Lipsitz, A.-H.S. Breizat, E.P. Dellinger, T. Herbosa, S. Joseph, P.L. Kibatala, M.C.M. Lapitan, A.F. Merry, K. Moorthy, R.F. Reznick, B. Taylor, and B.B. Gawande. A surgical safety check- list to reduce morbidity and mortality in a global population. *NEJM*, 360:491–499, 2009.

[101] J. Nascon and R.A. Schmidt. The activity-set hypothesis for warm-up decrement. *Journal of Motor Behavior*, 3:1–16, 1971.

[102] C.A. Wrisberg, A.W. Salmoni, and R.A. Schmidt. Warm-up effects in the learning of discrete motor skills. *Acta Psychologica*, 29:311–320, 1975.

[103] M.A. Anshel. The effect of arousal on warm-up decrement. *Research Quarterly for Exercise and Sport*, 56:1–9, 1985.

[104] M.H. Anshel and C.A. Wrisberg. Reducing warm-up decrement in the performance of the tennis serve. *J Sport and Exercise Psychology*, 15:290–303, 1993.

[105] P.A. Wetter. Resubmission of gap analysis workshop for training for reintegration of surgical skills. *Proceedings of the 2011 Society of Laparoendoscopic Surgeons Conference*, 2011.

[106] W. Arthur Jr., B. Winston Jr., P.L. Stanush, and T.L. McNelly. Factors that influence skill decay and retention: A quantitative review and analysis. *Human Performance*, 11:57–101, 1998.

[107] N.E. Seymour, A.G. Gallagher, S.A. Roman, M.K. OBrien, V.K. Bansal, D.K. Andersen, and R.M. Satava. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Ann Surg*, 236:458– 463, 2002.

[108] K. Kahol, R. Satava, J. Ferrara, and M.L. Smith. Effect of short-term pretrial practice on surgical proficiency in simulated environments: a randomized trial of the preoperative warm-up effect. *JACS*, 208:255, 2009.

[109] D. Calatayud, S. Arora, R. Aggarwal, I. Kruglikova, S. Schulze, P. Funch-Jensen, and T. Grantcharov. Warm-up in a virtual

Seattle Children's — Internal

reality environment improves perfor- mance in the operating room. *Ann Surg*, 251:1181, 2010.

[110] T.S. Lendvay, T.C. Brand, L. White, T. Kowalewski, S. Jonnadula, L. Mercer, D. Khorsand, J. Andros, B. Hannaford, and R.M. Satava. Virtual reality robotics surgery warm-up improves surgical performance: A prospective randomized con- trolled study. *JACS*, 216:1181–1192, 2013.

[111] D. Stefanidis, J.R. Korndorffer Jr., R. Sierra, C. Touchard, J.B. Dunne, and D.J. Scott. Skill retention following proficiency-based laparoscopic simulator training. *Surgery*, 138:165–170, 2005.

[112] Inc. Mimic Technologies. da vinci skills simulator (dvss), 2018.

[113] Y. Sharon, T.S. Lendvay, and I. Nisky. Instrument orientation metrics for robot- assisted and open surgical skill evaluation. *IEEE Transactions on Human-Machine Systems*, 2017.

[114] A.J. Hung, P. Zehnder, M.B. Patil, J. Cai, C.K. Ng, M. Aron, I.S. Gill, and M.M. Desai. Face, content and construct validity of a novel robotic surgery simulator. *Journal of Urology*, 186:1019–1025, 2011.

[115] J.B. Bakdash and L.R. Marusich. Repeated measures correlation, 2017.

[116] J.D. Kelly, M. Nash, N. Heller, T.S. Lendvay, and T.M. Kowalewski. Temporal variability of surgical technical skill perception in real robotic surgery. *Interna- tional Journal of Computer Assisted Radiology and Surgery*, SUBMITTED.

[117] N. Geri, A. Winer, and B. Zaks. Challenging the six-minute myth of online video lectures: Can interactivity expand the attention of learners? *Online Journal of Applied Knowledge Management*, 5(1):101–111, 2017.

[118] M. Schmitt, D.C.A. Bulterman, and P.S. Cesar. The contrast effect: Qoe of mixed-video qualities at the same time. *Quality and User Experience*, 3(7), 2018.

[119] Inc. C-SATS, 2020.

[120] H.W.R. Schreuder. Virtual reality training for robotic surgery. In *Hospital Health- care Europe*. Cogora Limited, 2014.

[121] T.S. Lendvay, T. Brand, A. French, L. Meryman, N. Organ, M. Sorensen, and T.M. Kowalewski. *Society of Pediatric Urology, Fall Congress*, 2018.

[122] A. Gawande. The number of surgical procedures in an american lifetime in 3 states. *Journal of the American College of Surgery*, 5(7):1688–1703, 1985.

[123] A. Zia, Y. Sharma, V. Bettadapura, E.L. Sarin, T. Ploetz, M.A. Clements, and I. Essa. Automated video-based assessment of surgical skills for training and eval- uation in medical schools. *International Journal of Computer Assisted Radiology and Surgery*, 11(9):1623–1636, September 2016.

[124] Y. Sharma, T. Plotz, N. Hammerld, S. Mellor, R. McNaney, P. Olivier, S. Desh- mukh, A. McCaskie, and I. Essa. Automated surgical osats prediction from videos. In *2014 IEEE 11th International Symposium on Biomedical Imaging – ISBI*, Bei- jing, China, April 29–May 2 2014. IEEE Xplore.

[125] A. Peuna, J. Hekkala, M. Haapea, J. Podlipska, A. Guermazi, S. Saarakkala, M. Nieminen, and E. Lammentausta. Variable angle gray level co-occurrence matrix analysis of t2 relaxation time maps reveals degenerative changes of carti- lage in knee osteoarthritis: Oulu knee osteoarthritis study. *Journal of Magnetic Resonance Imaging*, 47:1316–1327, 2018.

[126] G. Zhenhua, L. Zhang, and D. Zhang. A completed modeling of local binary pat- tern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010.

[127] J. Shi, X. Weng, S. He, and Y. Jiang. Biological motion cues trigger reflexive attentional orienting. *Cognition*, 117(3):348–354, 2010.

[128] S. Gharghabi, Y. Ding, C.M. Yeh, K. Kamgar L. Ulanova, and E. Keogh. Matrix profile viii domain agnostic online semantic segmentation at superhuman per- formance levels. In *2017 IEEE International Conference on Data Mining*, New Orleans, LA, Nov 18–Nov 21 2017. IEEE Xplore.

Appendix A

Robotic Surgery Readiness Supplemental Figures

The following supplemental figures are added, pertaining to Chapter 6, from the RSR study. Figure A.1 is a robotic surgery video from the Si platform after being edited to remove personally identifiable information of the performing surgeon. Figure
A.2 are plots of the kinematic tooltip positions from a sample video of robotic surgery. A screenshot of the website created to be used for obtaining faculty surgeon video reviews is shown in Figure A.3. Figure A.4 displays cumulative data from several different tasks combined of both the 3-D coordinates as well as position histograms in each coordinate plane, to give an idea for the general movements occurring during robotic surgery tasks.

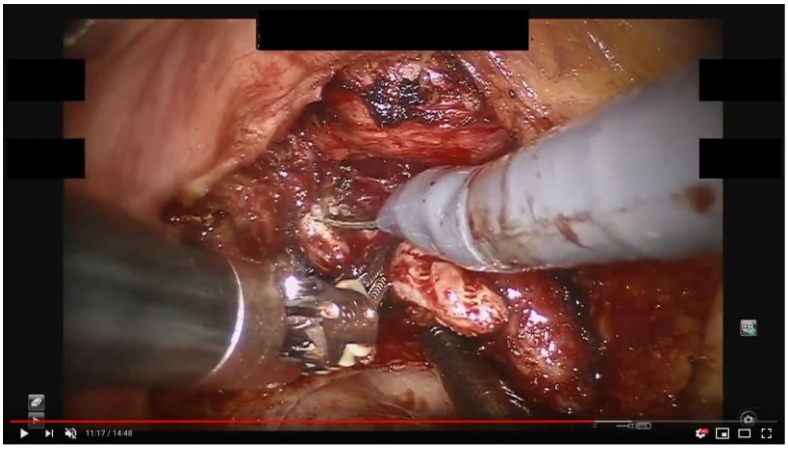


Figure A.1: Video of an Si daVinci surgery, after editing the video to the correct times and placing black boxes over the identifying pieces of information.

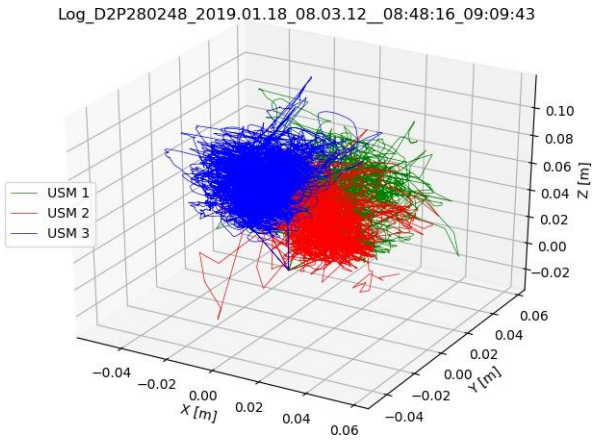


Figure A.2: 3-D Plot of the kinematic tooltips from a robotic surgery which corresponds with video edited at the same time range. USM 1-3 refers to the Universal Serial Manipulators, the surgical arms.



Figure A.3: A view of the website faculty surgeons would see when visiting to evaluate robotic surgeon technical skill.

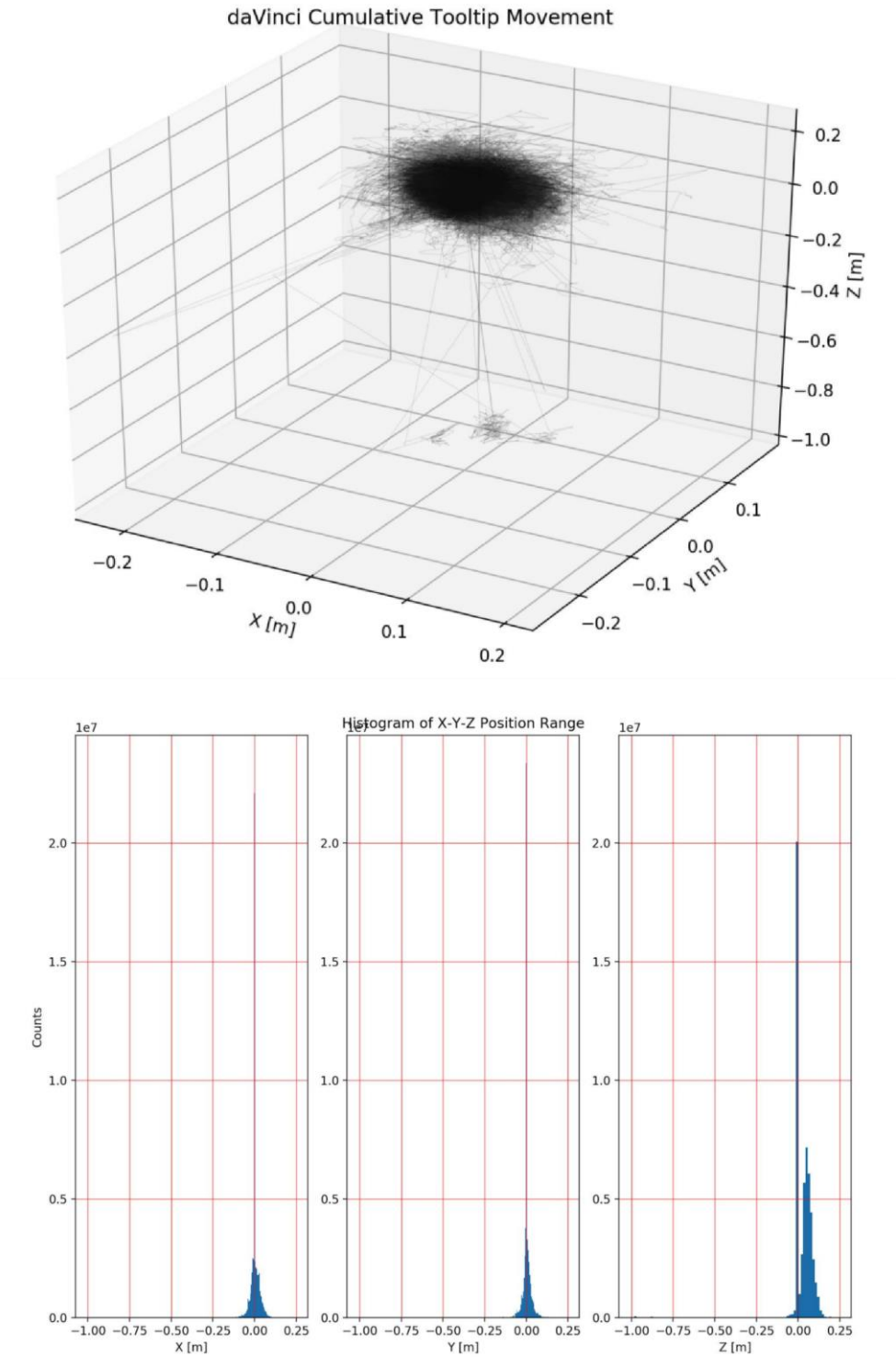


Figure A.4: Cumulative positions of several combined tasks from the RSR study, in both 3-D coordinates, as well as a histogram for frequency of position in each axis.

D.1 Kinematic Metrics

Several performance metrics were used to attempt to rank and visualize the relative skill of each performance. Some of these included rate of orientation change, normalized angular displacement [114], as well as spectral arc length (Chapter 4). Normalized angular displacement is defined as

$$\frac{1}{P}A = \sum_{j=1}^{N-1} |\delta\theta_{j,j+1}| \quad \text{(A.1)}$$

in which P is the path length, or the length of movement during a continuous movement between stops. This metric is a score of the overall change in orientation during a movement, which could also be thought of as the angular path. The rate of orientation change is

$$C = \frac{1}{\delta\theta_{j,j+1}} \quad \text{(A.2)}$$

$$N-1$$

$$N-1 \quad \delta t_{j,j+1}$$

$$i=1$$

where $t_{j,j+1}$ is the time difference between samples. Other computed metrics were mean velocity, median velocity, maximum velocity, minimum velocity, total time, path length, economy of motion, and acceleration.

A.0.2 Event Metrics

The da Vinci software records various event measures which have been shown in the past to correlate to skill. Of these, the metrics recorded for each case are shown in Table A.1.

Event	Description
ENERGY CTRL	Surgeon consoles are switched
CAMCTRL BTN	Camera is moved
ARM SWAP	Arms are switched between one another
HEAD SENSE	Surgeon presses head in or out of console
PORT CLUTCH	Surgeon presses clutch to move instruments

Table A.1: Recorded event measures for the RSR study.

.