The Relationship Between Technical Skills, Cognitive Workload, and Errors During Robotic Surgical Exercises

Sidney I. Roberts, BA,1 Steven Y. Cen, PhD,2 Jessica H. Nguyen, BS,1 Laura C. Perez, MD,1 Luis G. Medina, MD,1 Runzhuo Ma, MD,1 Sandra Marshall, PhD,3 Rafal Kocielnik, PhD,4 Anima Anandkumar, PhD,4,5 and Andrew J. Hung, MD1,i

Abstract

**Purpose:** We attempt to understand the relationship between surgeon technical skills, cognitive workload, and errors during a simulated robotic dissection task.

**Materials and Methods:** Participant surgeons performed a robotic surgery dissection exercise. Participants were grouped based on surgical experience. Technical skills were evaluated utilizing the validated Global Evaluative Assessment of Robotic Skills (GEARS) assessment tool. The dissection task was evaluated for errors during active dissection or passive retraction maneuvers. We quantified cognitive workload of surgeon participants as an index of cognitive activity (ICA), derived from task-evoked pupillary response metrics; ICA ranged 0 to 1, with 1 representing maximum ICA. Generalized estimating equation (GEE) was used for all modelings to establish relationships between surgeon technical skills, cognitive workload, and errors.

**Results:** We found a strong association between technical skills as measured by multiple GEARS domains (depth perception, force sensitivity, and robotic control) and passive errors, with higher GEARS scores associated with a lower relative risk of errors (all \( p < 0.01 \)). For novice surgeons, as average GEARS scores increased, the average estimated ICA decreased. In contrast, as average GEARS increased for expert surgeons, the average estimated ICA increased. When exhibiting optimal technical skill (maximal GEARS scores), novices and experts reached a similar range of ICA scores (ICA: 0.47 and 0.42, respectively).

**Conclusions:** This study found that there is an optimal cognitive workload level for surgeons of all experience levels during our robotic surgical exercise. Select technical skill domains were strong predictors of errors. Future research will explore whether an ideal cognitive workload range truly optimizes surgical training and reduces surgical errors.

**Keywords:** technical skills, cognitive workload, surgeon performance, surgical errors, robotics

Introduction

Robotics are an expanding technology that is already prevalent and continues to increase in use across several fields including urology. Surgical performance and skill in robotic surgery have been shown to account for as much as 25% of variation in patient outcomes.1-3 Evaluation of robotic surgical performance with technical skill metrics is thus an important component of improving overall surgeon performance and outcomes for patients.

Success of surgery is often measured by functional outcomes or surgical complications. However, surgical outcomes can also be measured by discrete errors committed during an operation.4,5 Discrete errors committed during surgery may not cause any significant complications but can be a surrogate for significant consequences in the future, hence leading to more consequential surgical complications.6 Previous study has shown that novice and expert surgeons can be distinguished based on error patterns.7 To our knowledge, no studies have specifically assessed the relationship of surgeon technical skills based on validated metrics and surgical errors.

To reduce the number of errors and improve outcomes, it is necessary to consider surgeon performance. A variety of evaluation tools and metrics have been introduced to break
down surgeon performance during robotic surgery. Technical skills are currently manually documented metrics that allow for formative feedback. To date, several skills-focused evaluation tools of surgeon performance have been validated.

Global Evaluative Assessment of Robotic Skills (GEARS) has been previously externally validated by our group and others, showing that surgeons with higher GEARS scores during robot-assisted radical prostatectomy (RARP) also had greater efficiency.8–10 Furthermore, technical skill evaluation was a robust predictor of urinary continence recovery after RARP.11 For the mentioned reasons, this study focuses on technical skills as a measure of surgeon performance.

Cognitive workload, or mental strain in the working memory, can influence surgeon performance and patient outcomes, and has been shown to differ between surgeon experience levels.12,13 For example, in our previous study, at the same cognitive workload level, novices and experts displayed significantly different surgical efficiency.14 Task-evoked pupillary response (TEPR) is based on changes in pupillary dilation/constriction and has previously been shown to be an indicator of human cognitive processing and workload.15,16

Our study seeks to elucidate the relationship between surgeon technical skills, cognitive workload, and errors during a robotic dissection task in a training environment.

Materials and Methods

Participants

This study was a retrospective cohort study using data obtained from a previously described study.14 Participants were selected to create three groups based on robotic surgical expertise: novice (no prior surgical experience), intermediate (<100 robotic cases), and expert (≥100 cases). The study included faculty surgeons, robotic fellows, urology residents, and medical students at our institution. Categorization of experience was based upon our prior published study on surgical performance.8

Technical skills

Technical skills were evaluated utilizing the GEARS assessment tool.8 The GEARS assessment included five metrics of performance: depth perception, bimanual dexterity, efficiency, force sensitivity, and robotic control. Two video raters received standardized training by the senior author (A.J.H) and proceeded to grade all participants independently. The raters were blinded to the participants’ level of experience. Interobserver reliability between two assessors was estimated using intraclass coefficients (ICCs). The two raters achieved an ICC value of >0.60 for overall GEARS scores, which is comparable to prior studies reporting GEARS interobserver variability.8–10

Determination of surgical errors

All participants underwent the same standardized robotic surgery dissection exercise (Fig. 1A). Participants were instructed to remove the top of a clementine orange in a premarked circular pattern. They then proceeded to dissect out a single clementine wedge. They were instructed to dissect as carefully as possible without tearing the skin of the orange or puncturing the “parenchyma,” or fruit, of the clementine. Upon retrospective review, errors were recorded with time stamps manually. The raters were blinded to the participant’s level of experience. Errors were categorized as either active if the surgeon commits injury to the fruit during active dissection movements (i.e., active cutting, tissue spreading, or skin peeling) or passive if the error occurred during supporting movements (i.e., tissue retraction).

Active errors: An “active puncture” occurred when the participant punctured the inner tissue of the clementine with an actively dissecting arm. In addition, an “active skin tear” occurred when participants tore the skin while attempting to dissect and remove it off the fruit’s inner tissues. For example, an active error may occur if the participant is using the right arm to separate two planes of tissue, but inadvertently punctures the tissue with the dissecting right arm.

Passive errors: A “passive skin puncture” occurred when participants drove an arm through the clementine outer skin during retraction, and a “passive fruit puncture” where participants passively punctured the fruit’s inner tissue with a supporting or nondominant arm. As an example, a passive error may occur if the participant is actively dissecting the
clementine fruit with the lefthand instrument, but retracting tissue with the righthand instrument, which inadvertently punctures the tissue (Fig. 1B).

Cognitive workload

We quantified cognitive workload of participants utilizing TEPR. TEPR was recorded through an eye-tracking device (Tobii Pro Glasses 2, Tobii Technology, Inc.) that measured changes in pupillary dilation at a rate of 100 Hz. The data were analyzed by EyeTracking, Inc.’s proprietary EyeWorks® software. The output of the analysis was the index of cognitive activity (ICA), ranging from 0 to 1. This represented the surgeon’s cognitive workload, with values closer to 1 representing higher cognitive processing and workload. ICA scores were calculated every second throughout the duration of the exercise. ICA scores were then averaged for each participant over the duration of the task.

Statistical analysis

Generalized estimating equation (GEE) was used for all modelings since it is flexible with data distribution and hierarchical structure. Identity link function with Gaussian distribution was used for continuous data. For count data (error count/100 second), we used a log link function with negative binomial distribution. Since some of our data (e.g., ICA and error count (recorded by second)) were clustered within each participant, we used a hierarchical modeling approach with GEE estimation.

Compound symmetry assumption was used for within-subject correlation structure. Model integrity was examined using residual plots. Dispersion assumption was tested by Pearson chi-square test and dispersion coefficient. For all models, the ratio between Pearson chi-square and degree of freedom is close to 1 with the testing p-value >0.05, indicating a well-fitted model without over dispersion. When exploring the association between individual GEARS subdomains and various error types, the Benjamini Hochberg procedure was used to control for multiple comparison error. SAS 9.4 was used for all data analyses.

Results

Participants

Upon review, 22 participant videos of the 27 from our original study were available. Data from five participants were lost because of technical issues (i.e., data corruption). Of the available subjects, seven were novices, nine were intermediates, and six were experts (Table 1).

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Novice</th>
<th>Intermediate</th>
<th>Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of participants</td>
<td>7</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Age</td>
<td>26.0 (23.5–31.5)</td>
<td>35.0 (32.5–40.5)</td>
<td>42.5 (33.25–47.5)</td>
</tr>
<tr>
<td>Years of practice</td>
<td>0.0 (0.0–5.5)</td>
<td>6.0 (3.0–7.5)</td>
<td>14.0 (6.25–24.0)</td>
</tr>
<tr>
<td>Number of robotic cases</td>
<td>0.0 (0.0–0.0)</td>
<td>50.0 (22.5–80.0)</td>
<td>525.0 (175.0–2500.0)</td>
</tr>
</tbody>
</table>

Table 1. Demographics by Experience Group

Surgeon experience vs technical skills

When comparing technical skills as measured by GEARS across the experience categories, expert surgeons had the highest overall total GEARS score of 96% ± 6%, intermediates had the second highest of 91% ± 7%, and novices had the lowest of 83% ± 13% (p = 0.02).

Surgeon experience vs errors

When comparing the overall error rates (passive and active combined) across the three experience categories, novices and intermediates made more overall errors than experts; however, these comparisons were not statistically significant. Novices had an error rate of 1.11 errors/100 seconds, whereas intermediates had a rate of 1.43 errors/100 seconds, and experts had a rate of 0.7 errors/100 seconds (p = 0.38). Novices had the highest rate of passive errors of 0.23 errors/100 seconds, whereas intermediates had a rate of 0.19 errors/100 second and experts had a rate of 0.05 errors/100 seconds (p = 0.23).

Technical skills vs errors

We found a strong association between technical skills, as measured by the five individual domains of GEARS, and errors made during the dissection exercise (Fig. 2). After controlling for the inflated error from multiple comparisons, the remaining adjusted statistically significant associations were depth perception and passive skin puncture with a relative risk (RR) of 0.28, 95% confidence interval (CI): (0.14 to 0.57) p < 0.01; between force sensitivity and passive fruit puncture with an RR of 0.20, 95% CI: (0.10 to 0.39) p < 0.01; between robotic control and passive skin puncture with an RR of 0.42 95% CI: (0.25 to 0.71) p < 0.01.

After controlling for possible confounders including cognitive workload (as measured by average max ICA) and surgeon experience, technical skills still had consistent and robust predictive effects on surgical errors. Level of surgeon experience was not a significant predictor of errors during the dissection exercise (Table 2).

The RR represented a log-linear trend, interpreted as a ratio of error rate when quality index increased by a unit. For example, when GEARS score for robotic control increased 1 U from 4 to 5 (perfect), the passive skin puncture error rate decreased exponentially from 0.09 errors/100 seconds to 0.04 errors/100 seconds with a log linear slope of 0.42 (Fig. 3).

Cognitive workload vs technical skills

Overall, throughout the dissection exercise, novices had an average ICA of 0.5, whereas experts and intermediates had an
average ICA of 0.41 and 0.43, respectively (p = 0.08). When comparing novices with intermediates and experts combined, novices had a higher average ICA (p = 0.03) (Fig. 4).

The correlation between technical skills using GEARS and average ICA was then assessed (Fig. 2). For novices, as average GEARS scores increased, the average ICA decreased, with the maximal novice GEARS scores occurring at an ICA level of 0.47 (95% CI: 0.42 to 0.53). Similarly, as GEARS scores increased for intermediates, the average ICA decreased, with the maximal GEARS scores occurring at an ICA level of 0.41 (95% CI: 0.37 to 0.46).

In contrast, as average GEARS score increased for experts, the average ICA increased, with the maximal GEARS scores occurring at an ICA level of 0.42 (95% CI: 0.32 to 0.52). When technical skills were poor as measured by GEARS, novices exhibited a high ICA level of 0.53 (95% CI: 0.45 to 0.56). In contrast, when projected technical skills were poor, experts exhibited a low ICA level of 0.11 (95% CI: −0.06 to 0.28). At the maximal GEARS score level, all three groups reached a similar ICA range of 0.41 to 0.47 (Fig. 5).

**Cognitive workload vs errors**

There were log-linear increases in the trends between the average ICA and corresponding error rates in novices with a slope of 1.22 (95% CI: 1.1 to 1.36, p < 0.01) and in intermediates with a slope of 1.09 (95% CI: 0.98 to 1.21, p = 0.1), but a decreasing trend in experts with a slope of 0.96 (95% CI: 0.84 to 1.11). When technical skills were poor as measured by GEARS, novices exhibited a high ICA level of 0.53 (95% CI: 0.45 to 0.56). In contrast, when projected technical skills were poor, experts exhibited a low ICA level of 0.11 (95% CI: −0.06 to 0.28). At the maximal GEARS score level, all three groups reached a similar ICA range of 0.41 to 0.47 (Fig. 5).

**TABLE 2. Predictors of Passive Surgical Errors**

<table>
<thead>
<tr>
<th>Predictor: technical skills (GEARS)</th>
<th>Outcome: passive surgical errors</th>
<th>Covariate (confounder)</th>
<th>Risk ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth perception Passive skin puncture</td>
<td>None</td>
<td>Average max ICA</td>
<td>0.21 (95% CI: 0.08–0.55)</td>
</tr>
<tr>
<td>Robot control</td>
<td>Passive fruit puncture</td>
<td>None</td>
<td>0.42 (95% CI: 0.25–0.71)</td>
</tr>
<tr>
<td>Force sensitivity</td>
<td>Passive fruit puncture</td>
<td>None</td>
<td>0.20 (95% CI: 0.1–0.39)</td>
</tr>
</tbody>
</table>

Technical skills as measured by GEARS scores were protective of passive errors such as passive skin punctures and passive fruit punctures in terms of relative risk. After controlling for possible confounders including cognitive workload (as measured by average max ICA) and surgeon experience, technical skills still had a consistent and robust protective effect on surgical errors.

CI = confidence interval; GEARS = Global Evaluative Assessment of Robotic Skills; ICA = index of cognitive activity.
CI: 0.84 to 1.09, \( p = 0.51 \), with a statistically significant interaction of \( p = 0.02 \). The log-linear trend was interpreted as a ratio of the error rate when ICA increased by a unit. For example, when ICA increased by 0.1 from 0.4 to 0.5, for novices, the overall error rate increased from 0.85 to 1.04 counts/100 seconds, with a ratio of 1.22. For experts, the error rate decreased from 0.7 to 0.67 with a ratio of 0.96 (Fig. 6).

Discussion

This study delineated the relationships between technical skills, cognitive workload, and errors during a simulated surgical task (Fig. 2). The primary outcome of our study design was surgical errors made during this dissection task. The strongest predictor of error occurrence was technical skills as measured by GEARS. We also found important associations between technical skills and cognitive workload based on different surgeon experience levels. As technical skills increased for novice and intermediates, their cognitive workload decreased. Conversely, as experts increased their technical skill level, workload increased to match a level similar to the ideal workload of novice and intermediates.

We found a similar trend when we assessed the relationship between cognitive workload and errors, with novices and intermediates committing more errors with increasing ICA, whereas experts committed less errors with increasing ICA. The importance of our findings is that there seems to be an optimal cognitive workload for surgeons regardless of

FIG. 3. Risk of errors during dissection exercises based on GEARS domain grades. Shown is the relationship of GEARS domains and the relative risk of passive errors such as passive fruit punctures and passive skin punctures. As the score for each individual domain improved for participants, the relative risk of passive errors significantly decreased. GEARS = Global Evaluative Assessment of Robotic Skills. Color images are available online.
experience, and that improved technical skills occur at different cognitive workload levels based on experience, potentially explaining why some errors may occur during surgery.

This study was able to provide further validation of the GEARS assessment tool of technical skills, as higher GEARS scores were associated with decreased errors during the surgical task. Higher GEARS scores were in fact a more significant predictor of error occurrence than surgical experience alone, indicating that regardless of experience, optimal surgeon technical skills are still essential for reducing error rates. We were able to show that when surgeons perform at lower technical skills, they are more likely to commit errors. Importantly, experience level was not a significant predictor of reduced errors upon multivariate analysis.

We also assessed cognitive workload as measured by ICA across the three experience groups. When assessing average ICA during the dissection exercise, novices were found to have significantly higher ICA values than intermediates and experts. We also found a correlation between technical skills and cognitive workload that differed between the groups as already noted. Thus, we were able to connect all three of these metrics of surgeon performance in this study. As novices and intermediates performed better as measured by their technical skills, their average ICA decreased.

Perhaps when novices were overstimulated and their cognitive workload was higher than average, they were more likely to make errors. In our previous study, we found that as cognitive workload increases, novices are more likely to increase their relative instrument speed, whereas experts were more likely to decrease their relative instrument speed. Similarly, with higher cognitive workload, novices exhibited lower technical skill, and through the association we found between technical skills and errors, this lower technical skill may make them more prone to errors. Experts, in contrast, were found to have relatively increased cognitive workload with increased technical skills. Thus, it is possible that when experts commit errors, they may be understimulated and not at their optimal cognitive workload.

**FIG. 4.** Average ICA based on expertise level. Novices had the highest average cognitive workload throughout dissection exercises; however, this was not significantly different when compared to intermediates and experts. The difference in ICA between novices and a combined group of intermediates and experts was statistically significant \((p < 0.05)\). ICA = index of cognitive activity. Color images are available online.

**FIG. 5.** Trends of ICA based on average GEARS scores for the three experience groups. *GEARS scores were tallied for each individual metric (i.e., Force Sensitivity, Robotic Control, etc.) and summed into a total score (maximum score of 25 across all metrics). The GEARS total was then calculated as a percentage of the total possible score. As total GEARS scores increased, novices and intermediates displayed a downward trend in cognitive workload as depicted by maximal ICA values. As total GEARS scores increased for experts, cognitive workload showed an upward trend. The best overall technical skills occurred at a similar cognitive workload level for all experience groups (0.41–0.47). Color images are available online.
Regardless of the task, experience is not only built through repetition and development of semiautomatic skills, but also a developed knowledge base that influences what experts notice, and how they organize, represent, and interpret information. Because of a lack of developed automaticity of movements, novices may be diverting a higher level of focus and cognitive workload to the specific movements required of the task, diverting attention away from other factors that are important for reduction of errors. Conversely as a result of increased experience, experts likely work at a lower cognitive workload at baseline because of developed automaticity, and instead have increased awareness of other aspects of the surgical task at hand, such as alternative courses of action and anticipation of adverse events.

This automaticity could explain a lower baseline cognitive workload level during the task for experts but could also leave them prone to errors if not enough cognitive workload or focus is given to the specific task at hand. Notably, the peak cognitive workload for experts, which occurred at the highest technical skills level, was in the same range as the cognitive workload for novices and intermediates at their highest technical skills level. Again, surgeons may have a common optimal cognitive workload regardless of their experience.

Our study had a few limitations. The errors that were assigned during this dissection exercise were determined retrospectively after anonymous review of the participant videos. This study is limited because of the small sample size; external validation of our results should be completed. Another limitation is the clinical relevance of the clementine dissection simulation exercise. Although the dissection of an orange does not necessarily equate to the difficulty or intensity of a live operation, the exercise is safe and able to mimic certain aspects of tissue dissection that surgeons routinely perform. The model used in our study was intended to simulate tissue dissection. There may be limitations of how outcomes of this model translate to specific procedures.

To our knowledge, this is the first study that has been able to associate surgeon performance, cognitive workload, and surgical errors with meaningful significance. Although this study assessed a robotic dissection exercise, the relationships identified should be applicable regardless of surgical approach (i.e., open vs robotic). Broadly, if we can identify why surgeons of different experience levels are making errors, we may be able to tailor our training to reduce errors during surgery.

Furthermore, if we can identify when surgeons are operating at suboptimal cognitive workload levels, we can potentially intervene to reduce errors. Based on our results, there does seem to be a common cognitive workload level where surgeons of all experience levels perform optimally. Future research should look to identify how this optimal cognitive workload can be maintained throughout surgery.

![Graph showing trends of total error count based on ICA for the three experience groups.](image-url)
Conclusion
This study found that select technical skill domains were strong predictors of errors during a robotic surgical dissection exercise. Furthermore, we found that there is an optimal cognitive workload level for surgeons of all experience levels. Future research can look to explore this optimal cognitive workload range to optimize surgical training and reduce surgical errors.

IRB Approval and Human and Animal Rights
Our study complied with protocols was approved by the University of Southern California’s IRB. 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. This article does not contain any studies with animals performed by any of the authors.

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

Authors’ Contributions
S.I.R. was in charge of project development, data collection and preparation, and article writing and editing. S.Y.C. performed data analysis and article writing and editing. J.H.N. performed article editing and project management. L.C.P. performed data collection and preparation and article editing. L.G.M. performed data collection and article editing. R.M. performed data collection, data analysis, and article editing. S.M. performed data analysis and article editing. R.K. performed data analysis and article editing. A.A. performed data analysis and article editing. A.J.H. was in charge of project development, data management, and article writing and editing.

Author Disclosure Statement
A.J.H. is a consultant for Mimic, Quantagene, and Johnson & Johnson. The study was not funded by any of these companies. Other authors have no conflict of interest.

Funding Information
Research reported in this publication was supported in part by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award No. K23EB026493, and by the National Cancer Institute under Award No. 1 R01CA251579-01A1.

References