Development and Validation of Objective Performance Metrics for Robot-Assisted Radical Prostatectomy: A Pilot Study

Andrew J. Hung,* Jian Chen, Anthony Jarc,† David Hatcher, Hooman Djaladat and Inderbir S. Gill

From the Center for Robotic Simulation & Education, Catherine & Joseph Aresty Department of Urology, University of Southern California Institute of Urology, Los Angeles, California (AJH, JC, DH, HD, ISG), and Medical Research, Intuitive Surgical, Inc., Norcross, Georgia (AJ)

Purpose: We explore and validate objective surgeon performance metrics using a novel recorder (“dVLogger”) to directly capture surgeon manipulations on the da Vinci® Surgical System. We present the initial construct and concurrent validation study of objective metrics during preselected steps of robot-assisted radical prostatectomy.

Materials and Methods: Kinematic and events data were recorded for expert (100 or more cases) and novice (less than 100 cases) surgeons performing bladder mobilization, seminal vesicle dissection, anterior vesicourethral anastomosis and right pelvic lymphadenectomy. Expert/novice metrics were compared using mixed effect statistical modeling (construct validation). Expert reviewers blindly rated seminal vesicle dissection and anterior vesicourethral anastomosis using GEARS (Global Evaluative Assessment of Robotic Skills). Intraclass correlation measured inter-rater variability. Objective metrics were correlated to corresponding GEARS metrics using Spearman’s test (concurrent validation).

Results: The performance of 10 experts (mean 810 cases, range 100 to 2,000) and 10 novices (mean 35 cases, range 5 to 80) was evaluated in 100 robot-assisted radical prostatectomy cases. For construct validation the experts completed operative steps faster ($p < 0.001$) with less instrument travel distance ($p < 0.01$), less aggregate instrument idle time ($p < 0.001$), shorter camera path length ($p < 0.001$) and more frequent camera movements ($p < 0.03$). Experts had a greater ratio of dominant-to-nondominant instrument path distance for all steps ($p < 0.04$) except anterior vesicourethral anastomosis. For concurrent validation the median experience of 3 expert reviewers was 300 cases (range 200 to 500). Intraclass correlation among reviewers was 0.6-0.7. For anterior vesicourethral anastomosis and seminal vesicle dissection, kinematic metrics had low associations with GEARS metrics.

Conclusions: Objective metrics revealed experts to be more efficient and directed during preselected steps of robot-assisted radical prostatectomy. Objective metrics had limited associations to GEARS. These findings lay the foundation for developing standardized metrics for surgeon training and assessment.

Key Words: robotic surgical procedures, prostatic neoplasms, education

Mounting evidence supports that surgical performance directly affects clinical outcomes and complication rates. Even after controlling for surgical volume and experience, for radical prostatectomy the differences
among surgeons can result in significant variation in quality of life and oncologic outcomes.3,4 One study found that during the initial diffusion from 2005 to 2007, robot-assisted radical prostatectomy was associated with an increased adjusted risk of any Patient Safety Indicators compared to open radical prostatectomy.5

Although robotic surgery has become an integral aspect of urological surgery, competency standards have not been set and proficiency evaluation varies widely among institutions.6 Typically the number of procedures performed or console hours are used as a surrogate measure for technical proficiency, despite the lack of sufficient evidence that these are accurate indicators of proficiency. Studies examining the learning curve for robotic surgeries vary widely in the number of cases to reach proficiency.7–9 An alternate measure of surgeon competence is performance based. Unfortunately, few credentialing programs are available to determine surgeon competence and standardized curricula for robotic surgery are still in development.10

Currently almost all surgical evaluative methods such as OSATS (Objective Structured Assessment of Technical Skills),11 GOALS (Global Operative Assessment of Laparoscopic Skills)12 and GEARS13 are based on direct observation by expert surgeons. More recent tools such as Prostatectomy Assessment and Competency Evaluation and RARP Assessment Score attempt to better provide standardized and procedure specific assessment.14,15 While some of these methods have been comprehensively validated and adopted to assess surgical performance, they are subjective, can vary greatly from different experts16 and are time-consuming.17 In 2007 approximately 100,000 prostatectomies were performed on the da Vinci Surgical System.18 By 2015 this annual number had increased to 145,000.19 If properly developed, truly objective, automated evaluation would provide a sustainable model for real-time feedback.

Prior studies have explored robotic surgical performance metrics in dry and wet laboratory settings, focusing mostly on basic metrics such as time to completion, instrument path length and error occurrence.20–22 Similarly, many validation studies for virtual reality simulation have examined computer generated performance metrics.23–25 Few of these metrics have been validated in the live clinical setting.

We captured objective surgeon performance metrics during actual, live RARP surgical cases to assess the difference in expert vs novice performance (construct validation) and associations of these objective metrics to subjective expert evaluation of performance (GEARS, concurrent validation). To our knowledge this is the first study to measure and analyze truly objective performance data directly from the robot during live surgery.

MATERIALS AND METHODS

Under institutional review board approved protocols for collection and analysis of surgical performance and patient clinical data, we collected synchronized video and objective surgeon performance data during 100 RARP cases on the da Vinci Si Surgical System using the da Vinci Logger (dVLogger, Intuitive Surgical, Inc.), a custom recording tool lent to us by Intuitive Surgical’s medical research team for the purposes of this study (fig. 1). The dVLogger captures synchronized video as endoscope view at 30 frames per second. Serial digital interface cables connect the left and right video output channels from the Vision Tower to the inputs on the dVLogger. System data (including kinematic data and events) are recorded at 50 Hz using an ethernet connection to the back of the Vision Tower. Kinematic data included characteristics of movement such as instrument travel time, path length and velocity. Events included frequency of master controller clutch use, camera movements, third arm swap and energy use.

Work Flow

All system data were recorded in an encrypted format along with synchronized video. Following our published graduated learning curriculum for robotic surgery, each RARP case was segmented into 13 discrete steps.26 A single operator (JC) prospectively annotated each video indicating the start/stop times for each step of the procedure and the performing surgeon (determined by attending surgeon in each case). These annotations directed which segments of the corresponding system data were to be analyzed. Any step performed by multiple surgeons was excluded from analysis. For this study we preselected the 4 steps of RARP performed by the widest range of surgeon experience, namely bladder mobilization, seminal vesicle dissection, anterior aspect of vesicourethral anastomosis and right standard pelvic lymph node dissection. The anonymized system data were sent to Intuitive’s medical research team with the segmentation...
start/stop times for decryption and then sent back to our institution for statistical analysis.

**Participants**
Participants in the study were urological residents, robotics fellows and faculty surgeons. Based on learning curve studies showing that patient outcomes plateau after a surgeon performs 100 RARP cases, training surgeons were defined a priori as having performed less than 100 robotic console cases at the start of the study and expert surgeons were defined as having performed 100 or more cases.

**Statistical Analysis**
Statistical analysis was done using IBM® SPSS® 24 with p < 0.05 considered statistically significant. The median and 95% CI were used to report performance metrics. For the 4 study RARP steps we compared the objective metrics between the training surgeons and experts (construct validation) using multilevel mixed effect modeling where repeat performances by the same surgeons were accounted for. When pertinent, this modeling also adjusted for performance metrics for patient body mass index, prostate volume and prostate specific antigen. Cohen’s d effect size was calculated to quantitatively measure the strength of difference between expert and novice performance. Cohen’s d >0.8 was considered a large effect size.

Correlation of objective metrics and GEARS measured concurrent validity. Three expert robotic surgeons (more than 100 cases) independently and blindly reviewed the first 5 minutes of video footage of the SVD and AA of each case using GEARS. In consultation with C-SATS, Inc. (Seattle, Washington), experts in soliciting surgical video evaluations suggested that raters would not likely have the attention span to review beyond a 5-minute video clip for assessment. Reviewers were excluded from rating cases they performed. Inter-observer reliability among raters was estimated using ICC. Spearman’s correlation test was used to evaluate the associations between objective performance metrics and expert evaluation.

**RESULTS**

**Demographics**
A total of 20 surgeons participated in the study, including 10 experts (mean 810 console cases, range 100 to 2,000) and 10 novices (mean 35 cases, range 5 to 80) (table 1). Expert surgeons included 7 attending surgeons and 3 clinical fellows. Novices included 1 attending surgeon, 2 fellows and 7 residents. Surgeon performance of 100 RARP cases from August 2016 to March 2017 at our main teaching hospital was evaluated. All 100 cases were managed with pelvic lymph node dissection (standard template 52%, extended 48%). No intra-operative complications or blood transfusions were noted.

**Learning Curve**
Scatter plots of surgeon performance metrics as a function of prior surgeon robotic experience show consistent trends (fig. 2). With increasing experience each of the 4 study steps took less time to complete (fig. 2, a). With experience a surgeon’s efficiency in instrument path also improved in all steps of RARP studied (fig. 2, b). Similarly, dominant hand instrument average velocity increased with surgeon experience (fig. 2, c). Average nondominant instrument velocity also increased with experience (data not shown). The number of cases needed for 50% or 75% improvement in task completion time, instrument path length and dominant hand velocity did not differ statistically among the 4 steps (p >0.05).

**Construct Validation**
Across the study steps consistent trends were observed between expert and novice performance (table 2). In terms of kinematic metrics the experts completed all 4 steps faster than novices (p <0.001) while travelling less distance to complete each step (p <0.01). Collective instrument idle time was also less in all steps (p <0.001) except during RLD (p >0.05). The dominant and nondominant hand instruments of experts spent less time moving (p <0.001), travelled less distance (p <0.004) and moved faster (p <0.05) than novices (except nondominant instrument in RLD). Expert camera path length was shorter during all 4 steps analyzed (p <0.05). Figure 3 illustrates the 3-D instrument path tracings of expert and novice examples during the 4 study steps. For event metrics in all steps the experts moved the camera more frequently than the novices (p <0.03). Experts used master controller clutching more during BM (p=0.003). Except for AA the experts swapped to the third arm more frequently (p <0.04) and experts used the energy pedal more frequently than novices (p <0.001).

| Table 1. Surgeon participant demographics |
|---------------------------------|---|---|---|
| Mean age (range)                | 42 (36–58) | 35 (30–42) | 0.03 |
| Mean prior robotic experience (range): | 7 (3–13) | 2 (0.5–4) | 0.003 |
| Total robotic cases             | 810 (100–2,000) | 35 (5–60) | 0.02 |
| RARP cases                      | 110 (50–410) | 18 (3–35) | 0.02 |
| No. training level:             |                               |   |   |
| Attending surgeons              | 7 | 1 |   |
| Urological fellows              | 3 | 2 |   |
| Urological residents            | 0 | 7 |   |
| No. surgeon hand dominance:     |   |   |   |
| Rt                               | 10 | 9 |   |
| Lt                               | 0 | 1 |   |

**Lymph Node Dissection**
On average the experts yielded more lymph nodes than novices during standard (p=0.01) and extended (p=0.046) template dissection. Experts also took less time per node (p <0.001).
Bimanual Dexterity
In all steps except AA the ratio of dominant-to-nondominant instrument path distance was greater with experts compared to novices (fig. 4, p < 0.04). Similarly, in all steps except AA the ratio of D:ND instrument velocity was greater with experts (p < 0.03). In all steps except SVD the ratio of D:ND instrument moving time was no different between expert and novice. During SVD the experts moved their dominant instrument over more time (p < 0.001).

Concurrent Validation
Median robotic experience of the 3 expert reviewers was 300 cases (range 200 to 500). As a measure of inter-rater reliability, ICC among reviewers examining total GEARS score was 0.6 (95% CI 0.41–0.71) for SV and 0.7 (95% CI 0.56–0.78) for AA. Table 3 summarizes the GEARS scoring between cohorts. For AA and SV the kinematic metrics had statistically significant but low associations with corresponding GEARS metrics (table 4). Little to no statistical association was observed between events metrics and GEARS metrics.

DISCUSSION
This study provides initial construct validation of kinematic and events metrics and limited concurrent validation of kinematic metrics captured during live RARP surgery. It marks the first time to our knowledge that truly objective metrics were collected and analyzed to assess live RARP surgeon performance. Our findings provide a foundation for understanding how to quantify robotic surgeons for training and assessment purposes.
### Table 2. Construct validation of kinematic metrics

<table>
<thead>
<tr>
<th></th>
<th>BM Expert (54 cases)</th>
<th>BM Novice (42 cases)</th>
<th>SVD Expert (75 cases)</th>
<th>SVD Novice (24 cases)</th>
<th>AA Expert (75 cases)</th>
<th>AA Novice (24 cases)</th>
<th>RLD Expert (31 cases)</th>
<th>RLD Novice (21 cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (95% CI)</td>
<td>p Value</td>
<td>Mean (95% CI)</td>
<td>p Value</td>
<td>Mean (95% CI)</td>
<td>p Value</td>
<td>Mean (95% CI)</td>
<td>p Value</td>
</tr>
<tr>
<td><strong>Kinematic metrics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task completion</td>
<td>10.6 (9.2–12.5)</td>
<td>&lt;0.001</td>
<td>8.7 (7.9–9.6)</td>
<td>&lt;0.001</td>
<td>15.5 (13.9–17.1)</td>
<td>&lt;0.001</td>
<td>1.4 (5.2–6.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Moving time (min)</td>
<td>25 (21.4–28.5)</td>
<td>&lt;0.001</td>
<td>45.2 (40.9–49.5)</td>
<td>&lt;0.001</td>
<td>14.6 (13.5–15.8)</td>
<td>&lt;0.001</td>
<td>1 (9.4–10.3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Idle time (sec)</td>
<td>29 (21–38)</td>
<td>&lt;0.001</td>
<td>246 (51–404)</td>
<td>&lt;0.001</td>
<td>17 (12–21)</td>
<td>&lt;0.001</td>
<td>11 (9–13)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Dominant instrument:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Path length (m)</td>
<td>15.8 (13.8–17.7)</td>
<td>&lt;0.001</td>
<td>28.2 (25.4–30.9)</td>
<td>&lt;0.001</td>
<td>12.4 (11.1–13.6)</td>
<td>&lt;0.003</td>
<td>0.8 (4.9–5.8)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Moving time (min)</td>
<td>8.8 (7.6–9.2)</td>
<td>&lt;0.001</td>
<td>21.9 (20.4–23.1)</td>
<td>&lt;0.001</td>
<td>7.4 (6.7–8)</td>
<td>&lt;0.001</td>
<td>1.2 (4.5–5.8)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Velocity (cm/sec)</td>
<td>2.9 (2.8–3)</td>
<td>&lt;0.001</td>
<td>2.1 (2–2.2)</td>
<td>&lt;0.001</td>
<td>2.2 (2.1–2.2)</td>
<td>&lt;0.001</td>
<td>1.8 (1.7–1.8)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Idle time (sec)</td>
<td>120 (84–170)</td>
<td>&lt;0.001</td>
<td>411 (257–755)</td>
<td>&lt;0.001</td>
<td>110 (91–119)</td>
<td>&lt;0.001</td>
<td>31 (28–42)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Nondominant instrument:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Path length (m)</td>
<td>8.1 (6.8–9.6)</td>
<td>&lt;0.001</td>
<td>18.2 (16.4–20)</td>
<td>&lt;0.001</td>
<td>1.3 (4–4.7)</td>
<td>&lt;0.001</td>
<td>0.8 (6–8)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Moving time (min)</td>
<td>9 (7.6–11.2)</td>
<td>&lt;0.001</td>
<td>24.1 (22–26.1)</td>
<td>&lt;0.001</td>
<td>8.2 (7.6–9.1)</td>
<td>&lt;0.001</td>
<td>14 (12.3–16.3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Velocity (cm/sec)</td>
<td>1.6 (1.5–1.6)</td>
<td>&lt;0.001</td>
<td>1 (1.2–1.3)</td>
<td>&lt;0.001</td>
<td>0.9 (0.9–1)</td>
<td>&lt;0.001</td>
<td>0.8 (0.7–0.8)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Idle time (sec)</td>
<td>91 (69–127)</td>
<td>&lt;0.001</td>
<td>392 (185–569)</td>
<td>&lt;0.001</td>
<td>42 (39–48)</td>
<td>&lt;0.001</td>
<td>127 (79–132)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Camera path length (m)</td>
<td>1.8 (1.4–2.3)</td>
<td>&lt;0.001</td>
<td>3.6 (2.9–4.2)</td>
<td>&lt;0.001</td>
<td>0.7 (0.6–0.8)</td>
<td>&lt;0.001</td>
<td>1 (0.9–1.1)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Systemic event metrics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Camera movement</td>
<td>8.4 (7.8–8.4)</td>
<td>&lt;0.001</td>
<td>4.8 (4.8–5.6)</td>
<td>&lt;0.001</td>
<td>4.3 (3–4.2)</td>
<td>&lt;0.001</td>
<td>5.1 (4.9–5.6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Master clutch</td>
<td>0.4 (0.3–0.5)</td>
<td>&lt;0.001</td>
<td>0.1 (0.1–0.2)</td>
<td>&lt;0.003</td>
<td>0.2 (0.1–0.2)</td>
<td>&lt;0.04</td>
<td>0.1 (0.1–0.2)</td>
<td>&lt;0.02</td>
</tr>
<tr>
<td>Third arm swap</td>
<td>2.4 (2.4–3)</td>
<td>&lt;0.001</td>
<td>1.8 (1.2–1.8)</td>
<td>0.03</td>
<td>5 (6–6.9)</td>
<td>5 (4.9–5.2)</td>
<td>1 (0.5–1.6)</td>
<td>0.2 (0.1–0.4)</td>
</tr>
<tr>
<td>Energy activation</td>
<td>10.8 (9.6–11.4)</td>
<td>&lt;0.001</td>
<td>7.2 (6.6–7.8)</td>
<td>&lt;0.001</td>
<td>7.2 (6–9)</td>
<td>4.6 (4.4–4.9)</td>
<td>9 (8.4–9.6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Lymph node dissection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nodes yielded</td>
<td>11 (9–14)</td>
<td>&lt;0.001</td>
<td>7 (6–8)</td>
<td>&lt;0.01</td>
<td>2.5 (1.9–3)</td>
<td>&lt;0.01</td>
<td>4 (3.4–4.6)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

BM results adjusted for body mass index, SVD results adjusted for prostate volume and prostate specific antigen, and AA results adjusted for prostate volume.
The surgeons (residents, fellows, attendings) involved in this study represent a wide range of experience, although in general the supervising attending surgeons are high volume surgeons. In this pilot study we were careful not to overanalyze the data in the learning curve plots (fig. 2). However, broad trends were seen. Experience did appear to improve efficiency. Through the kinematic and events metrics, a robust difference was detected in efficiency between the expert and novice groups as expected. Construct validation was conclusively established. However, certain findings were not as expected. The dominance of the dominant instrument was more pronounced in 3 of the study steps of RARP (all dissecting tasks). Perhaps unmeasured was the potentially increased efficiency that the expert surgeon brought, compared to the novice, in using the nondominant hand to support their dominant hand. Only in AA, a suturing task, did we not see a difference in the ratio of dominant-to-nondominant hand use between cohorts, suggesting that the AA step truly requires bimanual dexterity.

Among 3 expert surgeons reviewing the surgical videos the agreement was not statistically strong (ICC 0.6-0.7) despite the use of a widely validated assessment tool. While having a limited number of raters may have impacted the ICC, this finding is...
consistent with other reports and reaffirms the need for more objective methods to evaluate surgeons.\textsuperscript{13,16,27} In general we found weak relationships between kinematic metrics and corresponding GEARS metrics, and we did not find any association between events metrics and GEARS metrics. It is important to note that when there is no robust association between GEARS and objective metrics, there is simply not a correlation between them and it does not determine which is a more accurate measure of surgeon performance.

Where no statistical association was found, particularly with events metrics, it is possible that objective metrics such as frequency of clutch or camera use are able to capture aspects of technical performance more effectively than expert reviewers using GEARS.

Our study has limitations. Surgeon “style” was not accounted for, and may vary widely and impact surgical performance as measured simply by kinematic and events data. Likewise, many other patient variables that were not easily quantified may confound our results. That said, adjustment for patient variables did not significantly alter the results. The selected steps of RARP might not be representative of the entire surgery. However, we intentionally selected steps that training surgeons were allowed to perform under supervision to capture the broad learning curve of live surgery. Even so, there was a disproportionate number of experts performing each of the study steps compared to novices, which was simply a reflection of the practice at our institution. The metrics presented in this initial study are basic measures of surgeon performance, including instrument kinematics (eg time of movement, distance travelled and velocity) and console events (eg master clutch use, energy application and third arm swap). There is presently no measure of tissue handling or force sensitivity, which would arguably reflect surgeon performance more so than kinematic and events data. Concurrent validation assessment correlated objective metrics to a global

\textbf{Figure 4.} Ratio of nondominant-to-dominant hand use for instrument path distance, instrument moving time and average instrument velocity. Asterisk indicates $p <0.05$. During dissection tasks (BM, SVD, RLD) expert surgeons used dominant hand more compared to novices. There was no difference in degree of bimanual dexterity during AA.

\begin{table}[h]
\centering
\caption{GEARS score of experts and novices performing SVD and AA}
\begin{tabular}{lllll}
\hline
 & \textbf{Expert} & \textbf{Novice} & \textbf{p Value*} & \textbf{Cohen’s d} \\
 & Mean Score & Mean Score & & \\
 & (95\% CI) & (95\% CI) & & \\
\hline
\textbf{SVD:}\textsuperscript{†} & & & & \\
Depth perception & 3.8 (3.6–3.9) & 3.3 (3.1–3.6) & $<0.001$ & 0.9 \\
Bimanual dexterity & 3.8 (3.7–4.0) & 3.2 (2.9–3.4) & $<0.001$ & 1.1 \\
Efficiency & 3.5 (3.4–3.7) & 2.8 (2.6–3.1) & $<0.001$ & 0.9 \\
Force sensitivity & 3.6 (3.4–3.7) & 3.1 (2.8–3.4) & 0.02 & 0.7 \\
Robotics control & 3.9 (3.7–4.0) & 3.2 (3.0–3.4) & $<0.001$ & 1.3 \\
Overall score & 18.9 (18.3–19.5) & 15.7 (14.7–16.8) & $<0.001$ & 1.2 \\
\hline
\textbf{AA:}\textsuperscript{‡} & & & & \\
Depth perception & 4.0 (3.9–4.1) & 3.2 (3.0–3.5) & $<0.001$ & 1.4 \\
Bimanual dexterity & 4.0 (3.9–4.1) & 3.1 (2.9–3.4) & $<0.001$ & 1.3 \\
Efficiency & 3.8 (3.7–4.0) & 3.0 (2.8–3.2) & $<0.001$ & 1.3 \\
Force sensitivity & 4.0 (3.9–4.0) & 3.3 (3.0–3.5) & $<0.001$ & 1.4 \\
Robotics control & 4.1 (3.9–4.2) & 3.3 (3.1–3.5) & $<0.001$ & 1.6 \\
Overall score & 19.9 (19.3–20.5) & 15.9 (15.0–17.0) & $<0.001$ & 1.7 \\
\hline
\end{tabular}
\begin{flushleft}
* $p <0.05$ statistically significant.  \\
\textsuperscript{†} For 52 experts and 26 novices.  \\
\textsuperscript{‡} For 54 experts and 24 novices.
\end{flushleft}
\end{table}
assessment tool (GEARS) and not to a procedure specific assessment tool. That will be represented in future work. Finally, the small number of surgeons and cases evaluated in this pilot study limited the power of our conclusions. Nonetheless, certain statistical trends have been noted.

Future work in this field will delve into the more complex steps of RARP and other robotic surgeries, and more complex kinematic and events metrics will be explored and studied. With refined programming, automated objective metrics can be generated directly and immediately after an operation, simultaneously with minimal human processing. With sophisticated data analyses this feedback will be procedure and step specific to focus training.

CONCLUSION
Basic objective kinematic and events metrics can distinguish between expert and novice surgeons during live RARP (construct validation). Expert surgeons do not agree on what constitutes good surgery. Limited associations between objective metrics and subjective gold standard evaluation of surgeon performance leave room for further work to understand the applications and limitations of objective performance evaluation.

ACKNOWLEDGMENTS
Jie Cai and Christianne Lane provided statistical support, and Omid Mohareri provided dVLogger support. Intuitive Surgical provided dVLoggers to support the study.

REFERENCES
3. Cooperberg M, Odisho A and Carroll P: Outcomes for radical prostatectomy: is it the singer, the song, or both? J Clin Oncol 2012; 30: 476.