JOURNAL OF ENDOUROLOGY
Volume XX, Number XX, XXXXXX 2015

Pp. ■■■–■■■
DOI: 10.1089/end.2015.0665

Original Research

Crowdsourcing Assessment of Surgeon Dissection of Renal Artery and Vein During Robotic Partial Nephrectomy: A Novel Approach for Quantitative Assessment of Surgical Performance

Mary K. Powers, MD,[1] Aaron Boonjindasup, LMD,[1] Michael Pinsky, LMD,[1] Philip Dorsey, LMD,[1] Michael Maddox, LMD,[1] Li-Ming Su, MD,[2] Matthew Gettman, MD,[3] Chandru P. Sundaram, MD,[4] Erik P. Castle, MD,[5] Jason Y. Lee, MD,[6] and Benjamin R. Lee, MD[1]

Abstract

Introduction: We sought to describe a methodology of *crowdsourcing* for obtaining quantitative performance ratings of surgeons performing renal artery and vein dissection of robotic partial nephrectomy (RPN). We sought to compare assessment of technical performance obtained from the crowdsourcers with that of surgical content experts (CE). Our hypothesis is that the *crowd* can score performances of renal hilar dissection comparably to surgical CE using the Global Evaluative Assessment of Robotic Skills (GEARS).

Methods: A group of resident and attending robotic surgeons submitted a total of 14 video clips of RPN during hilar dissection. These videos were rated by both crowd and CE for technical skills performance using GEARS. A minimum of 3 CE and 30 Amazon Mechanical Turk crowdworkers evaluated each video with the GEARS scale.

Results: Within 13 days, we received ratings of all videos from all CE, and within 11.5 hours, we received 548 GEARS ratings from crowdworkers. Even though CE were exposed to a training module, internal consistency across videos of CE GEARS ratings remained low (ICC = 0.38). Despite this, we found that crowdworker GEARS ratings of videos were highly correlated with CE ratings at both the video level ($R = 0.82$, $p < 0.001$) and surgeon level ($R = 0.84$, $p < 0.001$). Similarly, crowdworker ratings of the renal artery dissection were highly correlated with expert assessments ($R = 0.83$, $p < 0.001$) for the unique surgery-specific assessment question.

Conclusions: We conclude that crowdsourced assessment of qualitative performance ratings may be an alternative and/or adjunct to surgical experts' ratings and would provide a rapid scalable solution to triage technical skills.

Introduction

Technical errors of surgeon skill can lead to patient morbidity and mortality, and assessment of skill can directly predict patient outcomes.[1] Malpractice claims involving surgical trainees have been shown to be caused by error in manual technique in as much as 56% of cases.[2] Surgical performance metrics include length of surgery, economy of motion, and mistakes/errors. In addition, the use of validated assessment tools, such as OSATS (Objective Structured Assessment of Technical Skills) or GEARS (Global Evaluative Assessment of Robotic Skills), provides objective feedback for surgical training.[1] Unfortunately, surgical skills assessment is not always standardized and can be expensive and time consuming. Assessments can take several days, if not weeks, for a small number of expert surgeons to review surgical footage and complete the evaluations.

Crowdsourcing relies on a large group of people, untrained in a specific field, to perform appraisal tasks. Crowdsourcing platforms, such as Amazon Mechanical Turk (AMT), allow

[1]Department of Urology, Tulane University School of Medicine, New Orleans, Louisiana.
[2]Department of Urology, University of Florida School of Medicine, Gainesville, Florida.
[3]Mayo Clinic, Rochester, Minnesota.
[4]Department of Urology, Indiana University School of Medicine, Indianapolis, Indiana.
[5]Department of Urology, Mayo Clinic, Scottsdale, Arizona.
[6]Department of Urology, St. Michael's Hospital University of Toronto, Toronto, Canada.
The abstract was presented at the World Congress of Endourology in October 2015 in London.

Global Evaluative Assessment of Robotic Skills (GEARS)*

Robotic Dissection of Renal Artery and Vein During Partial Nephrectomy

Trainee: **Date of Surgery:** **Evaluator:**

Please circle the number corresponding to the candidate's performance in each category, irrespective of training level.

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

	N/A	1	2	3	4	5
Dissection of Renal Artery:		Excessive force applied to vascular structures. Exposure limited.		Able to complete dissection safely with moderate guidance and telestration.		Optimal dissection and isolation of the renal artery with good exposure and dissection.

***RATE THE LEARNERS PERFORMANCE: circle the number that best describes their performance on this case:**

1: requires significant practice/needs improvement 2: good/at appropriate level 3: excellent/established

*Modified from rubric previously published: Goh et al. J of Urol. 2012

FIG. 1. Global Evaluative Assessment of Robotic Skills (GEARS) with novel renal artery dissection question.

anyone older than 18 years with access to the Internet to earn a supplemental income by reviewing audio or video clips and performing the tasks requested. Salaries earned by workers range from $0.05 to $50 per task. Protein-folding problems, medical diagnosis, and robotic skill testing in the dry lab have been some of the various scenarios posed to crowdworkers.[3,4] To date, crowdsourcing has not been used to evaluate surgical skill during robotic renal surgery in humans.

We sought to describe a methodology of crowdsourcing for obtaining quantitative performance ratings of surgeons performing procedures, specifically during renal artery and vein dissection of robotic partial nephrectomy (RPN). We sought to compare assessment of technical performance obtained from the crowdworkers with that of surgical content experts (CE). Our hypothesis is that the crowd can score performances of renal hilar dissection comparably to surgical CE using a validated robotic surgery assessment tool, as well as a novel renal artery dissection-specific skills question. In addition, performance scores can be generated at a much faster rate, with higher volumes of assessments being performed at a cost-effective rate.[5]

Methods

A group of robotic surgeons ($n=5$) ranging from resident to attending physician submitted a total of 14 video clips of RPN during hilar dissection from the time of renal vein exposure to renal artery isolation. All videos were 10 minutes in length and began at the time of renal vein identification with subsequent skeletonization of the hilar vasculature. Videos were selected in the third and fourth years of residency. They were selected not for the best performance but in a sequential temporal event during residency training of RPN. Videos chosen were used to demonstrate a wide range of experience so to see if one could discriminate the assessment of technique. Both CE and crowdworkers were blinded to the level of training.

These videos were rated by both crowd and CE for technical skills performance using the GEARS (Fig. 1). This clinical assessment tool has been validated and demonstrates excellent reliability for scoring for both resident and attending surgeons based on internal consistency (Cronbach's $\alpha=0.90$–0.93).[6] We additionally obtained a five-point Likert scale rating of a novel renal artery dissection question. Crowdworkers were recruited from AMT, while our surgical experts were recruited through e-mail and personal communication. The videos were assessed by both crowdworkers and CE through a service (Crowd-Sourced Assessment of Technical Skills; C-SATS, Inc., Seattle, WA) that merges surgical media with reviewers using validated assessment tools.[7]

Crowdworkers had to complete at least 100 Human Intelligence Tasks and obtain a 95% approval rating by Mechanical Turk, which were criteria described by Chen et al.[4] Crowdworkers were anonymous and could be identified by user ID codes.[3] CE were all urologic surgeons who specialized in robot-assisted laparoscopic surgery, particularly with experience in RPN, and had experience with a minimum of 500 robotic cases as primary surgeon.

In an effort to establish inter-rater reliability, CE completed a training module that allowed them to compare their scores with the average CE score on a set of videos that were previously assessed. The training module included four training videos along with a fifth test video that was used to determine whether the CE score would be included in the final assessment. Experts were required to score within three points of the previously established gold standard rating for the test video in an effort to optimize inter-rater reliability among the videos to be rated. Two of the five experts did not pass this criterion and were not included in subsequent ratings and analyses.

The videos were uploaded to a Web site that demonstrated five grading domains with a free response field. To ensure that workers were not clicking random responses, there were ''attention check'' questions that confirmed quality answers.[3] Based on Chen et al., an estimated 30 responses/video from workers would be necessary to obtain a 95% confidence interval to be ±1 point on the grading scale.[4]

A minimum of 3 CE and 30 AMT crowdworkers evaluated each video with the GEARS rating instrument. The crowdworkers were paid, while the CE were not. We evaluated inter-rater reliability of experts using a two-way intraclass

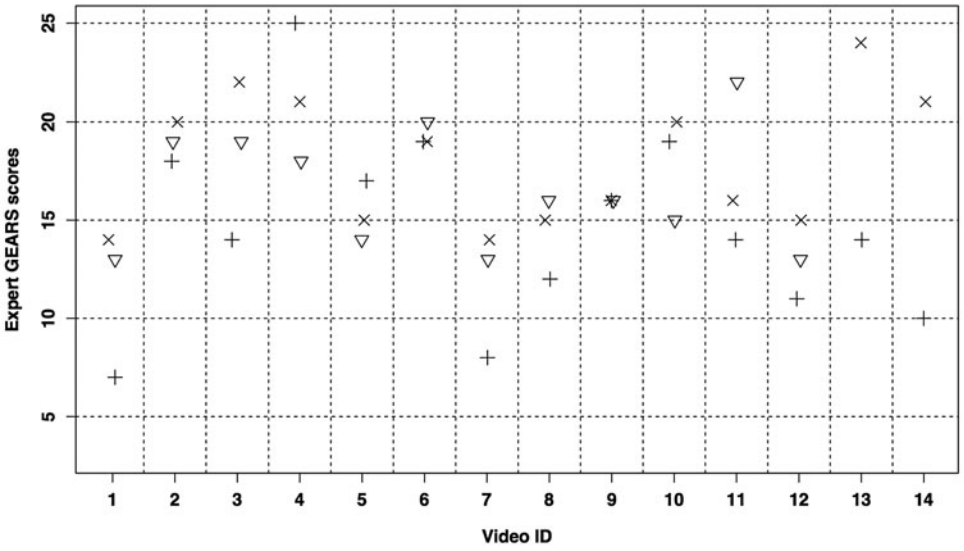

FIG. 2. Inter-rater variability of content expert (CE) GEARS scores across videos (ICC = 0.38).

correlation coefficient. We used linear mixed-effects models to derive an average crowd and CE rating for each video clip and compared rating modalities using graphical plots of the data and summarizing associations with Pearson correlation coefficients and linear regression models. We additionally examined surgeon-level GEARS scores by averaging scores.

Results

Within 13 days, we received ratings of all 14 videos from the expert surgeons compared to a time frame of 11 hours and 33 minutes from which we received 548 GEARS ratings from crowdworkers. This gave an average of 39 responses per video clip submitted. Even though CE were exposed to a training module, internal consistency across videos of CE GEARS ratings remained low (ICC = 0.38, Fig. 2). With the exception of force sensitivity (r = 0.35), all subdomains showed comparable correlations between the crowd and experts (0.76–0.83). This anomaly with force sensitivity may be worthy of additional investigation to examine whether this result is replicated in subsequent evaluations. Not all crowdworkers reviewed every video, so an ICC was not calculated. In addition, the large numbers within the crowd imply that there will not be an accurate measure of the true underlying quantity with good precision.

Despite this, we found that crowdworker GEARS ratings of videos were highly correlated with CE ratings at both the video level (R = 0.82, $p < 0.001$, Fig. 3) and surgeon level (R = 0.84, $p < 0.001$, Fig. 4), meaning there was agreement between the CE and crowdworkers. Video level *vs* surgeon level was stratified based on experience in surgical training. Similarly, crowdworker ratings of the renal artery dissection were highly correlated with expert assessments (R = 0.83, $p < 0.001$, Fig. 5) for the unique surgery-specific assessment question.

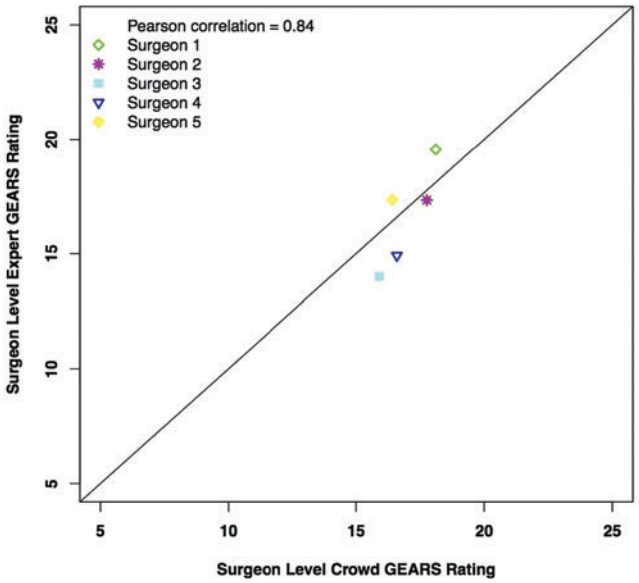

FIG. 4. Average surgeon-level GEARS ratings of CE *vs* crowdworkers. Color images available online at www.liebertpub.com/end

Discussion

Objective identification of technical skills is crucial to surgeon development, trainee feedback, and implementation of improvement processes. Historical beliefs have purported that proficiency of the operating surgeon is an integral factor in variation of clinical outcomes; however, objective measures to assess technical video analysis are lacking. Using surgeon experts to review surgical performance video clips can be time consuming as demonstrated in our study. In

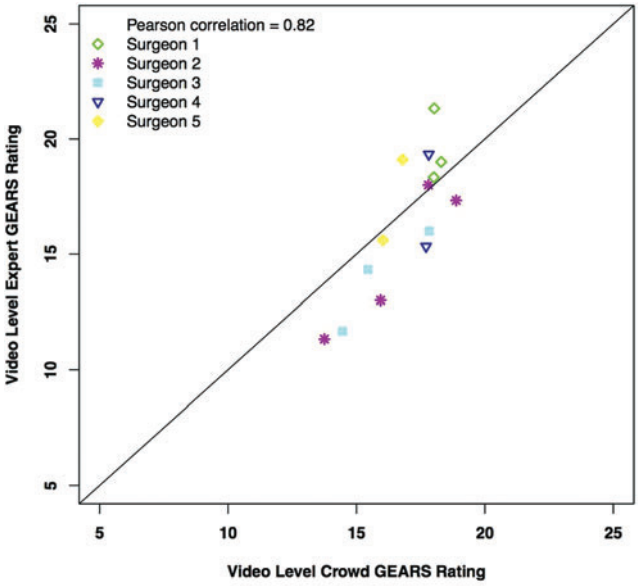

FIG. 3. CE GEARS ratings *vs* crowdworker GEARS ratings across videos. Color images available online at www.liebertpub.com/end

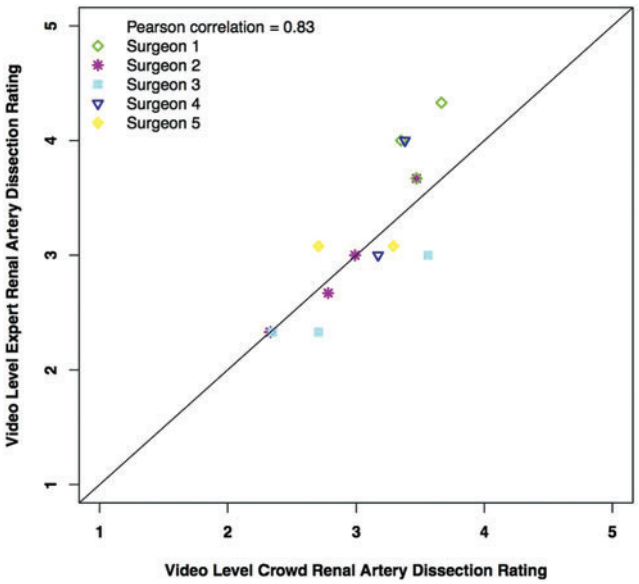

FIG. 5. Likert scale ratings of the renal artery dissection of CE *vs* crowdworkers. Color images available online at www.liebertpub.com/end

addition, there is a huge cost saving that can be accrued using crowdsourcing platforms, such as AMT. In addition, the use of crowdsourcing allows blinding of reviewees and anonymity of reviewers.

Lendvay et al. demonstrated that for three surgeon experts to review one 10-minute video would cost approximately $75–$126. In sharp contrast, workers get paid on average US $0.50–US $1.00 per video, markedly lower than the cost for experts.[3,8] C-SATS can help provide quick feedback and identify physicians who are deficient in particular areas of training.[7] This can help resident surgeons focus their studies on specific deficiencies without forcing all trainees to undergo broad-based curriculum training. Multiple studies have demonstrated the use of dry lab or porcine lab to provide feedback using crowdsourcing. Holst et al. have shown correlation using Cronbach's α statistic, with levels >0.9 indicating ''excellent agreement'' between CE and crowdworkers for robotic porcine bladder closure.[3,7] Six hundred crowdworkers evaluated all the video clips within 5 hours, while it took seven CE 14 days to complete evaluations.[3]

Pegboard transfer and laparoscopic suturing tasks have been used to demonstrate the ability of crowdworkers to accurately score surgeons. Forty-nine surgeons from novice to expert performed the above tasks, and three surgeon experts graded the tasks using the GEARS tool. This project determined similar success with high correlation between the crowd and surgeon scores, 0.84 for pegboard and 0.92 for suturing, computed using Cronbach's α.[5] Similarly, three surgeon experts contributed to scoring.

Facebook crowdworker has also been used by other authors to evaluate surgical proficiency. Chen et al. demonstrated that Mechanical Turk took 24 hours, Facebook took 25 days, and surgeon experts took 24 days to receive full response. The benefit of using Facebook crowdworkers was that no compensation was required.[4] Turnover time for surgeon experts is thought to be secondary to other job obligations and the extensive time demands that are required by reviewing videos.

C-SATS has been shown to correlate well with surgeon experts in human robotic surgery and could be used as an adjunct to first identify outliers for improvement, which would decrease the burden of applying the same training curriculum to all performers. This would allow for targeted training resource use aimed at improvement in the providers or trainees who would most benefit.[7]

Renal hilar dissection during RPN can be a challenging skill set to master, and objective feedback by crowdworkers can give similar ratings compared with surgeon experts. The limitation of this process is that there are qualitative characteristics of a renal hilar dissection, which can drive outcomes beyond just the technical skills score, such as exposure. We did not assess these because such factors are beyond the capture of the objective technical skills assessment tool. In addition, only 10-minute video clips were uploaded for review. It is possible that longer segments may generate different scores. However, based on past investigation, 10 minutes has been able to yield data between crowds and experts that correlate well. It is also possible that longer segments to review video may start to burden experts more and lead to decreased assessment participation.

In addition, only three surgeon experts were included in the results because two were eliminated due to criterion failure. This provides the limitation of fewer surgeons reviewing all the videos. Only 14 videos were reviewed, but on average, 39 reviews were done by crowdworkers. Future projects would allow for a large video sampling and varied robotic urologic procedures. Only five surgeons were used with 14 video clips, with the ability to expand upcoming studies to include a more volume.

The teaching process of robotic surgery can be a steep learning curve, where oftentimes, patient safety during critical portions of a procedure can require the attending surgeon to take control of a case. The main focus of our study was to evaluate varied skill-level surgeons during hilar dissection, but patient safety was always maintained during surgical dissection.

This was the first study to date objectively reviewing live human surgical procedures and scoring novice, intermediate, and expert surgeons compared with many previously evaluated dry lab or porcine surgical skills. This was a small cohort of video clips reviewed ($n = 14$), and further studies are needed to demonstrate persistent correlation of surgeon experts to crowdworker ratings.

Conclusion

We conclude that crowdsourced assessment of qualitative performance ratings of human surgical procedures may be a suitable alternative to surgical expert ratings. This would provide a rapid, cost-effective, scalable solution to triage technical skills and assessment among large groups of surgeons. Further validation of different types of surgeries will need to be validated to see if this method generalizes.

Acknowledgment

We would like to thank the staff at CSATS, Inc., Thomas Lendvay, Bryan Comstock, Justin Warren, Carstens Wisnes, Adam Monsen, and Sean O'Connor for their effort in orchestrating the assessment process.

Author Disclosure Statement

No competing financial interests exist.

References

1. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med 2013;369:1434–1442.
2. Rogers SO, Gawande AA, Kwaan M, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. Surgery 2006;140:25–33.
3. Holst D, Kowalewski TM, White LW, et al. Crowd-sourced assessment of technical skills: Differentiating animate surgical skill through the wisdom of crowds. J Endourol 2015;29: 1183–1188.
4. Chen C, White L, Kowalewski T, et al. Crowd-sourced assessment of technical skills: A novel method to evaluate surgical performance. J Surg Res 2014;187:65–71.
5. White L, Kowalewski T, Dockter R, et al. Crowd-sourced assessment of technical skills (C-SATS): A valid method for

discriminating basic robotic surgery skills. J Endourol 2015;29: 1295–1301.
6. Goh A, Goldfarb D, Sander J, et al. Global evaluative assessment of robotic skills: Validation a clinical assessment tool to measure robotic surgical skills. J Urol 2012;187:247–252.
7. Holst D, Kowalewski TM, White LW, et al. Crowd-sourced assessment of technical skills: An adjunct to urology resident surgical simulation training. J Endourol 2015;29:604–609.
8. Medscape 2013 Physician Compensation Report. 2013.

Address correspondence to:
Benjamin R. Lee, MD
Department of Urology
Tulane University School of Medicine
1430 Tulane Ave., SL-42
New Orleans, LA 70112-2632

E-mail: brlee@tulane.edu

Abbreviations Used

AMT = Amazon Mechanical Turk
CE = content experts
C-SATS = Crowd-Sourced Assessment of Technical Skills
GEARS = Global Evaluative Assessment of Robotic Skills
OSATS = Objective Structured Assessment of Technical Skills
RPN = robotic partial nephrectomy