available at www.sciencedirect.com
journal homepage: www.europeanurology.com

European Association of Urology

Brief Correspondence

Measuring to Improve: Peer and Crowd-sourced Assessments of Technical Skill with Robot-assisted Radical Prostatectomy

Khurshid R. Ghani[a,*], *David C. Miller*[a], *Susan Linsell*[a], *Andrew Brachulis*[a], *Brian Lane*[b], *Richard Sarle*[c], *Deepansh Dalela*[d], *Mani Menon*[d], *Bryan Comstock*[e], *Thomas S. Lendvay*[f], *James Montie*[a], *James O. Peabody*[d],
for the Michigan Urological Surgery Improvement Collaborative

[a] *Department of Urology, University of Michigan, Ann Arbor, MI, USA;* [b] *Department of Urology, Spectrum Health, Grand Rapids, MI, USA;* [c] *Michigan Institute of Urology, Dearborn, MI, USA;* [d] *Vattikuti Urology Institute, Henry Ford Health System, Detroit, MI, USA;* [e] *Department of Biostatistics, University of Washington, Seattle, WA, USA;* [f] *Department of Urology, University of Washington, Seattle, WA, USA*

Article info

Article history:
Accepted November 23, 2015

Associate Editor:
James Catto

Keywords:
Robotic surgery
Skills assessment
Crowd sourcing
Radical prostatectomy

Abstract

Because surgical skill may be a key determinant of patient outcomes, there is growing interest in skill assessment. In the Michigan Urological Surgery Improvement Collaborative (MUSIC), we assessed whether peer and crowd-sourced (ie, layperson) video review of robot-assisted radical prostatectomy (RARP) could distinguish technical skill among practicing surgeons. A total of 76 video clips from 12 MUSIC surgeons consisted of one of four parts of RARP and underwent blinded review by MUSIC peer surgeons and prequalified crowd-sourced reviewers. Videos were rated for global skill (Global Evaluation Assessment of Robotic Skills) and procedure-specific skill (Robotic Anastomosis and Competency Evaluation). We fit linear mixed-effects models to estimate mean peer and crowd ratings for each video. Individual video ratings were aggregated to calculate surgeon skill scores. Peers (n = 25) completed 351 video ratings over 15 d, whereas crowd-sourced reviewers (n = 680) completed 2990 video ratings in 38 h. Surgeon global skill scores ranged from 15.8 to 21.7 (peer) and from 19.2 to 20.9 (crowd). Peer and crowd ratings demonstrated strong correlation for both global (r = 0.78) and anastomosis (r = 0.74) skills. The two groups consistently agreed on the rank order of lower scoring surgeons, suggesting a potential role for crowd-sourced methodology in the assessment of surgical performance. Lack of patient outcomes is a limitation and forms the basis of future study.

Patient summary: We demonstrated the large-scale feasibility of assessing the technical skill of robotic surgeons and found that online crowd-sourced reviewers agreed with experts on the rank order of surgeons with the lowest technical skill scores.

* Corresponding author. Department of Urology, University of Michigan, North Campus Research Complex Building 16, 114W, 2800 Plymouth Road, Ann Arbor, MI 48109, USA. Tel. +1 734 615 4034; Fax: +1 734 232 2400.
E-mail address: kghani@med.umich.edu (K.R. Ghani).

Surgical performance is under increasing scrutiny from multiple stakeholders. Recent work has shown that among fully trained surgeons, technical skill correlates with patient outcomes [1]. For men with prostate cancer, outcomes of greatest importance after robot-assisted radical prostatectomy (RARP; ie, cancer control, continence, and potency) may depend on surgeon performance that may be discernable on video review. However, it has not been

http://dx.doi.org/10.1016/j.eururo.2015.11.028
0302-2838/Published by Elsevier B.V. on behalf of European Association of Urology.

established that the assessment of technical skill among practicing surgeons performing RARP is feasible with current instruments and technology. Furthermore, because peer assessment is time-consuming and expensive, there is a need to explore more scalable and reproducible strategies.

In this context, surgeons from the Michigan Urological Surgery Improvement Collaborative (MUSIC), a consortium of 42 urology practices comprising 85% of urologists in the state of Michigan [2], evaluated whether peer surgeon assessments of the technical quality of RARP were feasible. In addition, we assessed whether peer and crowd-sourced reviewers (*crowdworkers*; ie, anonymous lay reviewers from online communities [3]) could distinguish differences in technical skill among practicing surgeons.

All surgeons in MUSIC were invited to submit a representative video of nerve-sparing RARP. Videos were deidentified and edited by a quality coordinator into 76 video clips containing one of four parts of surgery: bladder neck dissection, apical dissection, nerve sparing, and urethrovesical anastomosis. Global robotic skills were assessed using the Global Evaluative Assessment of Robotic Skills (GEARS) instrument [4]. Videos of the complete unedited anastomosis were assessed using a procedure-specific instrument, the Robotic Anastomosis and Competency Evaluation (RACE) [5]. Finally, each video had a summary judgment question for overall skill in which the reviewer was asked to pass or fail the surgeon.

Individual video clips were evaluated by at least four peer reviewers from a total of 25 MUSIC surgeons. The process for crowd-sourced review was adopted from Chen et al [3], and reviews were obtained from prequalified crowdworkers using Amazon Mechanical Turk (Amazon.com Inc., Seattle WA, USA). Each video clip was evaluated by at least 30–55 crowdworkers. A detailed description of the video review and methods is provided in Supplementary Figures 1 and 2 and in Supplement 1.

Video-based assessments of technical skill were successfully completed by both groups of reviewers. Peers took 15 d to complete 318 global robotic skill and 33 anastomosis skill ratings. In comparison, crowdworkers completed 2531 global skill ratings within 21 h and 459 ratings of the anastomosis within 38 h. Global skill scores provided by peers had a wider range compared with those given by crowdworkers (Table 1) and varied across the 12 surgeons ($p < 0.001$). The interrater reliability among peers was higher for evaluations with RACE compared with GEARS (Krippendorff's $\alpha = 0.55$ and $\alpha = 0.25$, respectively). Case experience of the peer reviewer did not confer higher agreement of ratings.

Aggregate peer and crowd-sourced surgeon scores demonstrated a strong positive correlation for both global robotic (GEARS) (Fig. 1a) and anastomosis (RACE) (Fig. 1b) skills (Pearson correlation 0.78 and 0.74, respectively; $p < 0.001$). Importantly, both sets of reviewers agreed on the rank order of the lower scoring surgeons using both rating instruments (Table 1 and Supplementary Table 1). For the summary skill question, both groups agreed identically on the relative order of the passing rate for each surgeon (Supplementary Fig. 3). Notably, the lower three performing surgeons were the same three lowest performing surgeons with the global skills assessment. Supplementary Videos 1–4 show the nerve-sparing part of RARP by surgeons with high global skill scores from peers (Supplementary Video 1) and crowdworkers (Supplementary Video 2) and with low global skill scores from peers and crowdworkers (Supplementary Videos 3 and 4).

Our findings build on a recent landmark study demonstrating that the technical skill of practicing bariatric surgeons varied widely and correlated with postoperative outcomes [1]. Our study lays the foundations for the future assessment of the surgical skill of RARP in clinical practice. First, from a measurement perspective, we found that interrater agreement among peers improved when using a procedure-specific instrument. Although we evaluated only the anastomosis with RACE in a smaller cohort of 8 surgeons, our interrater reliability findings were comparable to the RACE validation study in which the instrument was tested on 28 surgeons with varying experience [5]. Lack of agreement among peer reviewers may reflect differences in training and experience. In addition, unlike Birkmeyer

Table 1 – Global robotic skill scores for surgeons evaluated for robotic prostatectomy by peer surgeons and crowd-sourced reviewers, sorted by peer rank

Surgeon ID	No. of peer reviewer ratings	Peer reviewer score, mean (95% CI)	Peer rank	No. of crowd reviewer ratings	Crowd reviewer score, mean (95% CI)	Crowd rank
1	30	21.7 (20.2–23.1)	1	231	20.9 (20.4–21.4)	5
2	26	21.0 (19.5–22.5)	2	201	20.3 (19.8–20.9)	7
3	21	20.4 (18.7–22.1)	3	174	20.7 (20.2–21.3)	6
4	24	20.5 (18.9–22.1)	4	200	20.9 (20.4–21.4)	4
5	17	20.5 (18.6–22.3)	5	132	21.8 (21.2–22.4)	1
6	24	19.4 (17.8–21.0)	6	207	21.2 (20.7–21.7)	2
7	29	19.2 (17.8–20.7)	7	236	20.9 (20.4–21.3)	3
8	20	18.8 (17.1–20.5)	8	170	20.0 (19.5–20.6)	9
9	30	18.4 (16.9–19.9)	9	228	20.2 (19.7–20.7)	8
10	29	18.2 (16.7–19.7)	10	227	19.9 (19.4–20.4)	10
11	31	16.2 (14.7–17.6)	11	236	19.5 (19.0–20.0)	11
12	37	15.8 (14.5–17.2)	12	289	19.2 (18.7–19.6)	12

CI = confidence interval; ID = identifier.
Mean values were calculated from a linear mixed-effects model using ratings across all video segments.

Please cite this article in press as: Ghani KR, et al. Measuring to Improve: Peer and Crowd-sourced Assessments of Technical Skill with Robot-assisted Radical Prostatectomy. Eur Urol (2016), http://dx.doi.org/10.1016/j.eururo.2015.11.028

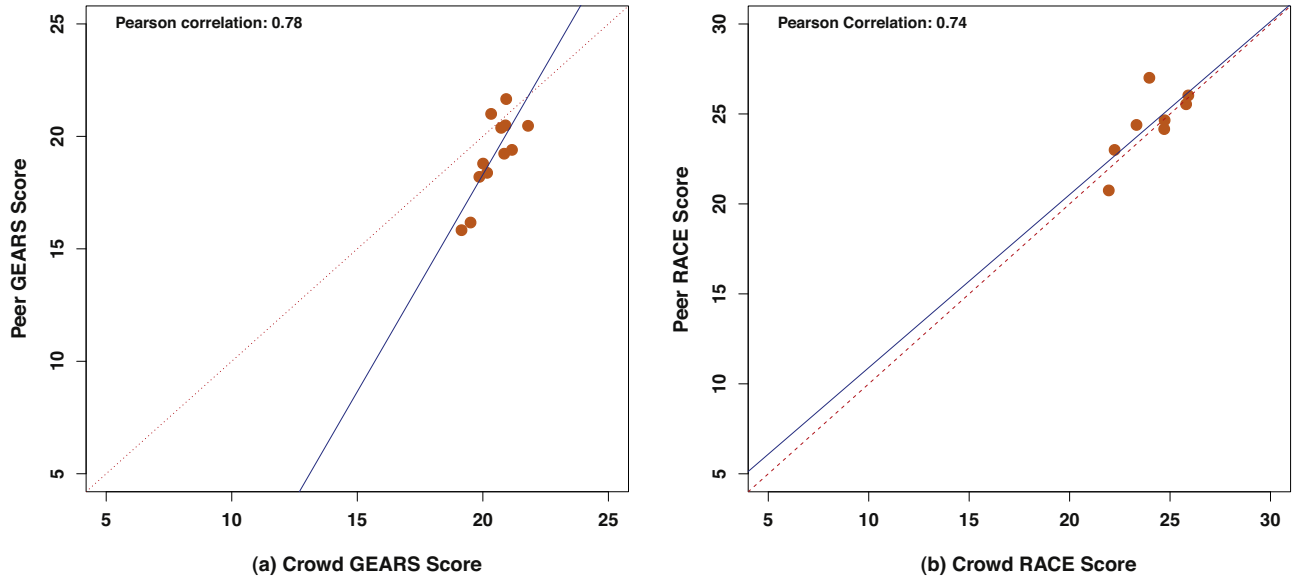

Fig. 1 – Correlation between peer and crowd-sourced technical skill scores for surgeons performing robot-assisted radical prostatectomy. (a) Global robotic skill (Global Evaluative Assessment of Robotic Skills) for 12 surgeons: Pearson correlation = 0.78; number of peer and crowd-sourced ratings per surgeon score ranged from 17 to 37 and 132 to 289, respectively. (b) Robotic anastomosis skill (Robotic Anastomosis and Competency Evaluation) for eight surgeons: Pearson correlation = 0.74; number of peer and crowd-sourced ratings per surgeon score was 4 and 55 to 60, respectively. GEARS = Global Evaluative Assessment of Robotic Skills; RACE = Robotic Anastomosis and Competency Evaluation.

and colleagues who distributed videos electronically to peer surgeons [1], we used a secure Web-based system that demonstrated feasibility for both peer and crowd-sourced review. This method is easier to administer at scale with the ability to maintain patient confidentiality.

Second, the ability of crowdworkers to provide rapid reviews suggests a potential role for crowd-sourced methodology in the evaluation of technical skills. Whereas peer review may have the greatest face validity, it is time consuming and costly to implement on a broad scale. Crowd-sourced assessment could serve as a filter through which lower performing surgeons could be identified for peer review and possibly for coaching initiatives. It remains to be established whether this methodology would be integrated into the evaluation of surgeons.

Our study has several limitations. Surgeon participation was voluntary, and video submission was of a representative nerve-sparing procedure. We also chose only four parts of the surgery for our evaluation. This was based on the template validated by Birkmeyer et al [1], who chose video segments of the most clinically significant and technically challenging portions of the case. Despite this, we demonstrated differences in skill even among this self-selected group. It is possible that in the future, surgeons might be willing to submit more difficult cases for review, and that could prove more useful for feedback and coaching. In addition, an instrument like GEARS was not designed specifically to assess RARP, and nuances in technique cannot be assessed through this tool [6]. Although the association between crowd and peer ratings was strong, the low agreement among peer reviewers, especially for GEARS, is a limitation. Moreover, although we used a 2-min video to introduce the instrument, we did not undertake standardized training for expert reviewers. Standardized training workshops to gather consensus could have improved agreement and should be considered in future studies. We did not do so because we wanted to compare untrained peer surgeons with untrained crowdworkers.

Furthermore, for reasons of safeguarding the anonymity of the submission process, we did not study the relationship between technical skill and case volume. In a post hoc preliminary analysis of surgeon skill score with perioperative morbidity, we found a weak positive correlation between skill and patient outcomes. However, our study has a limited sample of surgeons to assess this relationship adequately. Also, when averaging across large numbers of crowd-sourced responses, performances at the extreme ends of the technical rating scale will tend to be under- or overrated due to regression to the mean. The correlation with peer-based assessment will increase as the number of crowd-sourced ratings increases, although the correspondence with peer ratings may not be one to one. Programs wishing to use crowdworkers should be aware that although the rank order of technical performance will be preserved, crowd and peer ratings will not correspond exactly at the extremes.

Traditionally, measures such as complication rates and operative times have served as a surrogate for technical proficiency [7,8]. However, outcomes such as recurrence of the condition after surgery and patient quality of life—end points more likely linked to technical quality—cannot be derived from perioperative data [9]. Better skills may lead to improved patient care, which would ultimately benefit physicians, patients, and payers [10]. Moving forward, we intend to study the association between skill with outcomes using a larger sample of surgeons and longer follow-up. Moreover, our study identified the need for an improved observational instrument specific to RARP, and this task is in process.

Please cite this article in press as: Ghani KR, et al. Measuring to Improve: Peer and Crowd-sourced Assessments of Technical Skill with Robot-assisted Radical Prostatectomy. Eur Urol (2016), http://dx.doi.org/10.1016/j.eururo.2015.11.028

In conclusion, we demonstrated the large-scale feasibility of assessing the technical skill of practicing robotic surgeons. We found that both peer surgeons and layperson crowdworkers could identify differences in surgical skill with RARP. In addition, both groups consistently agreed on the rank order of surgeons with the lowest surgical skill scores across constructs and instruments, suggesting a potential role for crowd-sourced methodology in emerging quality improvement initiatives of surgical performance.

Author contributions: Khurshid R. Ghani had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Miller, Montie, Menon, Ghani, Peabody.
Acquisition of data: Ghani, Miller, Brachulis, Lendvay, Comstock, Dalela, Peabody.
Analysis and interpretation of data: Ghani, Miller, Lendvay, Comstock, Dalela, Peabody.
Drafting of the manuscript: Ghani, Miller.
Critical revision of the manuscript for important intellectual content: Ghani, Miller, Peabody, Dalela, Lendvay, Comstock, Lane, Sarle.
Statistical analysis: Comstock.
Obtaining funding: Miller, Montie, Linsell.
Administrative, technical, or material support: Linsell, Brachulis, Montie, Miller, Sarle, Lane.
Supervision: Miller, Peabody.
Other (specify): None.

Financial disclosures: Khurshid R. Ghani certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: Dr. Ghani receives contract support from Blue Cross Blue Shield of Michigan for serving as the co-director of the Michigan Urological Surgery Improvement Collaborative. Dr. Miller receives grant support from the National Cancer Institute; he also receives contract support from Blue Cross Blue Shield of Michigan for serving as the director of the Michigan Urological Surgery Improvement Collaborative. Thomas Lendvay and Bryan Comstock have equity interests in and are cofounders of C-SATS, Inc., a University of Washington start-up company. The other authors have nothing to disclose.

Funding/Support and role of the sponsor: Michigan Urological Surgery Improvement Collaborative is funded by Blue Cross Blue Shield of Michigan. The sponsor had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation or approval of the manuscript; and decision to submit the manuscript for publication.

Acknowledgments: The authors acknowledge the significant contribution of the clinical champions and urologists in each participating Michigan Urological Surgery Improvement Collaborative practice. In addition, we would like to acknowledge the support provided by David Share, Tom Leyden, Roz Darland, and the Value Partnerships program at Blue Cross Blue Shield of Michigan.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.eururo.2015.11.028.

References

[1] Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med 2013;369:1434–42.
[2] Montie JE, Linsell SM, Miller DC. Quality of care in urology and the Michigan Urological Surgery Improvement Collaborative. Urol Pract 2014;1:74–8.
[3] 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.
[4] 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. J Urol 2012;187:247–52.
[5] Raza SJ, Field E, Jay C, et al. Surgical competency for urethrovesical anastomosis during robot-assisted radical prostatectomy: development and validation of the robotic anastomosis competency evaluation. Urology 2015;85:27–32.
[6] Kowalczyk KJ, Huang AC, Hevelone ND, et al. Stepwise approach for nerve sparing without countertraction during robot-assisted radical prostatectomy: technique and outcomes. Eur Urol 2011;60:536–47.
[7] Begg CB, Riedel ER, Bach PB, et al. Variations in morbidity after radical prostatectomy. N Engl J Med 2002;346:1138–44.
[8] Finks JF, Osborne NH, Birkmeyer JD. Trends in hospital volume and operative mortality for high-risk surgery. N Engl J Med 2011;364:2128–37.
[9] Alderson D, Cromwell D. Publication of surgeon-specific outcomes. Br J Surg 2014;101:1335–7.
[10] Hampton T. Efforts seek to develop systematic ways to objectively assess surgeons' skills. JAMA 2015;313:782–4.