EUROPEAN UROLOGY 69 (2016) 547–550

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

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.

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

* 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).

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.