Wisdom of the crowds: Use of crowdsourcing to assess surgical skill of robot-assisted radical prostatectomy in a statewide surgical collaborative

James Peabody*, Detroit, MI, David Miller, Ann Arbor, MI, Brian Lane, Grand Rapids, MI, Richard Sarle, Dearborn, MI, Andrew Brachulis, Susan Linsell, Tae-Kyung Kim, Ann Arbor, MI, Deepansh Dalela, Mani Menon, Detroit, MI, James Montie, Ann Arbor, MI, Bryan Comstock, Tom Lendvay, Seattle, MI, Khurshid Ghani, Ann Arbor, MI

INTRODUCTION AND OBJECTIVES: Because a surgeon's technical skill may be a pivotal determinant of outcomes following robotic-assisted radical prostatectomy (RARP), the Michigan Urological Surgery Improvement Collaborative (MUSIC) examined the degree to which expert and crowd sourced reviewers can distinguish differences in performance among fully-trained urologists.

METHODS: MUSIC is a statewide consortium of 42 urology practices. Twelve MUSIC surgeons submitted a video of a representative RARP which were reviewed by 25 surgical experts from MUSIC and crowdsourced reviewers from the Amazon Mechanical Turk platform using the Crowd-Sourced Assessment of Technical Skills (C-SATS) on-line survey tool. Four key steps (bladder neck division, nerve sparing, apical dissection, and urethrovessical anastomosis) from each video were edited into 76 video-clips (maximum duration 10 minutes) for analysis using the Global Evaluation Assessment of Robotic Skills (GEARS) instrument. Unedited complete anastomosis clips from 8/12 surgeons (range 9-37 minutes) were also assessed using the Robotic Anastomosis and Competency Evaluation (RACE) instrument. Inter-rater reliability was evaluated using Krippendorff’s alpha and intraclass correlation coefficient (ICC). Linear mixed-effects models derived average crowd and expert ratings for each video-clip. Individual video-clip ratings were aggregated to provide surgeon scores.

RESULTS: MUSIC experts completed 318 video ratings in 15 days; crowd-workers completed 2,531 ratings in 21 hours. Expert GEARS ratings lacked sufficient internal consistency across videos (Krippendorff’s alpha=0.25, ICC=0.24) to validate against crowd-workers. The correlation between crowd and expert scores was higher (r=0.78) when using aggregated surgeon scores. For the RACE instrument, inter-rater reliability was better (ICC=0.55) and Pearson correlation with crowd scores was 0.74 (Figure 1). For both rating tools, experts and the crowd consistently agreed in the rank order of the lower scoring surgeons.

CONCLUSIONS: Expert surgeons and the crowd demonstrate high levels of agreement for identifying lower scoring RARP surgeons. The ability of crowds to provide rapid reviews of surgical videos suggests a potential role for crowdsourced methodology in emerging training and quality improvement initiatives of surgical performance.

Source of Funding: Blue Cross Blue Shield of Michigan

Correlation between expert and crowdsourced review of robot-assisted radical prostatectomy using (a) GEARS and (b) RACE instruments, aggregated by surgeon

[Graphs showing correlation between expert and crowd ratings]