Crowdsourcing in Surgical Skills Acquisition: A Developing Technology in Surgical Education

Jessica C. Dai, MD
Thomas S. Lendvay, MD
Mathew D. Sorensen, MD, MS

ABSTRACT

Background The application of crowdsourcing to surgical education is a recent phenomenon and adds to increasing demands on surgical residency training. The efficacy, range, and scope of this technology for surgical education remains incompletely defined.

Objective A systematic review was performed using the PubMed database of English-language literature on crowdsourced evaluation of surgical technical tasks up to April 2017.

Methods Articles were reviewed, abstracted, and analyzed, and were assessed for quality using the Medical Education Research Study Quality Instrument (MERSQI). Articles were evaluated with eligibility criteria for inclusion. Study information, performance task, subjects, evaluative standards, crowdworker compensation, time to response, and correlation between crowd and expert or standard evaluations were abstracted and analyzed.

Results Of 63 unique publications initially identified, 13 with MERSQI scores ranging from 10 to 13 (mean = 11.85) were included in the review. Overall, crowd and expert evaluations demonstrated good to excellent correlation across a wide range of tasks (Pearson's coefficient 0.59–0.95, Cronbach's alpha 0.32–0.92), with 1 exception being a study involving medical students. There was a wide range of reported interrater variability among experts. Nonexpert evaluation was consistently quicker than expert evaluation (ranging from 4.8 to 150.9 times faster), and was more cost effective.

Conclusions Crowdsourced feedback appears to be comparable to expert feedback and is cost effective and efficient. Further work is needed to increase consistency in expert evaluations, to explore sources of discrepant assessments between surgeons and crowds, and to identify optimal populations and novel applications for this technology.

Introduction

Traditional models of surgical training rely on the apprenticeship model and the Halstedian concept of graduated responsibility with advancement through residency.[1,2] The changing landscape of surgical training and practice, influenced by modern educational theory, new technologies, cost consciousness, work hour reform, and national patient safety concerns, necessitates a shift in the traditional paradigm of volume-based surgical competency for residents and practicing surgeons.[1–4] At the national level, malpractice claims have shown that 41% of errors in surgical care causing patient harm are technical, with operative skill itself having a direct correlation with surgical complications and patient outcomes.[5–7]

A critical component to mastering surgical technique is frequent, immediate feedback.[8,9] However, current feedback mechanisms remain limited in objectivity, timeliness, and scope. The Accreditation Council for Graduate Medical Education (ACGME) operative case logs are a surrogate for a trainee's surgical exposure and surgical skills, but may only accurately capture the true extent of resident involvement in 47% to 58% of cases.[10,11] Most trainees additionally receive informal feedback on their operative technique from surgical mentors, although the quality, quantity, and formative value of that feedback varies. A single surgeon's view may be biased and reflect only a limited breadth of observed surgical procedures.[12,13] Moreover, such feedback often is not timely. In 1 large academic orthopedic surgery program, 58% of residents reported that evaluations were rarely or never completed in a timely manner, with more than 30% completed more than 1 month after a rotation's end.[14]

To help standardize feedback, structured assessment tools have been developed, such as the Objective Structured Assessment of Technical Skills (OSATS), the Global Operative Assessment of Laparoscopic Skills (GOALS), and the Global Evaluative Assessment of Robotic Skills (GEARS; provided as online supplemental material).[15–17] While such metrics have been used by expert surgeons largely to evaluate videotaped simulation tasks and intraoperative surgery, this approach is not easily scalable. With the

DOI: http://dx.doi.org/10.4300/JGME-D-17-00322.1

Editor's Note: The online version of this study contains the surgical technical skills assessment instrument, the quality assessment scores of included studies, and the summary of evaluation metrics and agreement between crowds and experts.

growing use of simulation in surgical education, the resources required for video recording (in addition to the time and cost for each surgeon's participation) are significant.[18,19] This problem has been recognized nationally, and the Association for Surgical Education recently designated the determination of the best methods and metrics to assess surgical performance among its top 10 research priorities.[20]

In this context, interest in crowdsourcing has grown. Crowdsourcing refers to a problem-solving approach in which a specific task is completed more effectively by a large cohort of decentralized individuals than by any single person or small group.[21] Although participants may lack expertise within the relevant fields, the distributed wisdom of the group brings the advantages of efficiency, scalability, flexibility, and diversity to solving a particular problem.[22,23] The Internet has facilitated access to this technology: Amazon (Seattle, WA) Mechanical Turk is 1 example of an accessible online crowdsourcing platform.

Crowdsourcing has been successfully applied within medicine to help discover protein folding patterns, generate phylogenetic promoters, diagnose colonic polyps, and identify red blood cells infected with malaria.[24–27] In the clinical arena, it has been explored for diagnosing bladder cancer with confocal laser endomicroscopy, identifying diabetic retinopathy, and teaching at both the premedical and graduate medical education levels.[28–31] However, 1 of its most promising applications is in assessing technical skills, an area in particular need of innovation in the current training environment.

We performed a systematic review of the current literature about use of crowdsourcing technology in the evaluation of technical skills tasks to assess its efficacy, efficiency, and potential applicability across a wide range of surgical training contexts. We hypothesized that this technology is a valuable adjunct to traditional feedback mechanisms for technical skills development, both for trainees and experienced surgeons. Several areas of emerging research in the applications of crowdsourcing to surgical training are also explored.

Methods

A systematic literature search using PubMed (US National Library of Medicine, Bethesda, MD) was performed in February 2017. The search encompassed English-language articles using the Boolean search strings: "crowdsourcing" AND "surgery" (32 results); "crowdsourced assessment" AND "surgery" (3 results); "crowdsourcing" AND "technical skills" (14 results); and "crowdsourcing" AND "surgical skills" (14 results). Given the relative novelty of the subject, personal communication with subject matter experts was used to identify additional articles that may have been missed in the initial query. Results were combined and duplicates removed. Abstracts were screened by a single reviewer (J.C.D.), and nonrelevant publications were excluded (non-English articles, oral presentations, editorials, non–peer-reviewed publications, and articles using crowdsourcing for purposes other than evaluation of a surgical procedure or task performed by trainees or practicing surgeons).

The full texts of remaining articles were then reviewed to determine whether they met the following inclusion criteria: (1) peer-reviewed manuscript represented original research; (2) methodology and results were included; (3) crowdsourcing was used in an evaluative capacity; and (4) standardized metrics were used to evaluate task performance. References for those articles were reviewed to identify any additional articles that met inclusion criteria. The literature search was repeated on April 2, 2017, to ensure that no additionally published studies were missed. All studies were assessed for quality using the Medical Education Research Study Quality Instrument (MERSQI), and studies of the lowest quality (MERSQI score 5) were excluded.[32] Study background information (authors, year, journal, and methodology) was collected for each article. Data regarding performance task, subjects, evaluative standards, and crowdworker compensation were abstracted from each study, along with data on response times to queries and on the correlation between crowd and expert evaluations.

Results

Using the initial search criteria, 63 unique publications on crowdsourcing and technical skills evaluation were identified. Two articles identified in discussion with subject matter experts were already included in the literature search. Forty-eight abstracts were excluded for nonrelevant subject matter. Of the 15 full-text articles reviewed, 2 were excluded because they did not contain original research (1 clinical review article and 1 systematic review).[33,34] Review of references identified 1 additional publication, which was ultimately excluded because it was solely an abstract.[35] The remaining 13 articles were evaluated for quality, with MERSQI scores ranging from 10 to 13 and a mean MERSQI score of 11.85 (SD = 0.9; provided as online supplemental material). All 13 studies were well above the minimum MERSQI score for inclusion (FIGURE 1).

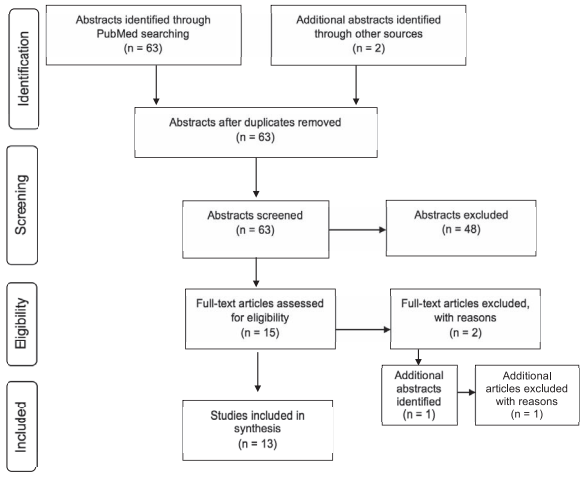

FIGURE 1
Flow Diagram for Study Inclusion and Exclusion in Systematic Review

The 13 studies assessed a wide array of technical tasks across a range of training levels (TABLE 1). Four evaluated laparoscopic simulation tasks exclusively,[19,36–38] 4 evaluated robotic simulation tasks exclusively,[39–42] and 2 evaluated procedural simulations.[43,44] Live intraoperative surgical performance was evaluated by 2 published articles.[45,46] Most studies evaluated a group of expert surgeons and residents across a range of experience levels. Of the 4 that did not, 1 examined a single "above average" performer,[39] 1 included only general surgery interns,[36] 1 assessed practicing urologists,[46] and 1 exclusively evaluated medical students.[47] All but 1 study used the Amazon Mechanical Turk platform to solicit nonexpert feedback; the remaining study recruited nonexpert respondents through a website,[41] and 1 study also queried inexperienced Facebook users for feedback.[39] Crowd and expert feedback was characterized using well-described, objective global evaluation metrics. Some studies additionally assessed individual task components. Two studies examining live surgical videos more specifically scored the performance of procedural subcomponents using unique task-specific evaluation score metrics.[45,46] A third study solicited crowd and expert preferences for more optimally performed segments of a single task.[41]

There appeared to be generally consistent correlations between crowds and experts for global technical performance ratings (provided as online supplemental material). Six studies reported that correlation for 16 unique tasks and 2 task subcomponents using Pearson's coefficients, with good overall correlation for global task performance (Pearson's $r = 0.75$–0.95) and component-specific scores (Pearson's $r = 0.74$–0.83).[36–38,42,45,46] The sole exception was a laparoscopic clip-applying task (Pearson's $r = 0.59$).[37] Six studies reported Cronbach's alpha scores to characterize the consistency between crowds and experts, with values greater than 0.9 indicating *excellent* agreement, 0.7 to 0.9 indicating *good* agreement, and below 0.5 indicating *poor* and *unacceptable* agreement.[19,38,40,43,44,47] For 10 tasks assessed in the studies, Cronbach's alpha ranged from 0.62 to 0.92, with a single outlier of $\alpha = 0.32$ in a cohort of medical students performing a fulguration exercise on a commercial virtual reality laparoscopic simulator (FIGURE 2).[47] The overall strong consistency was independent of the evaluation metric used. Where reported, mean score ratings between experts and crowds tended to be similar,[39,44,47] and there was good agreement in overall pass/fail decisions between crowds and experts.[36,38,46] Where assessed, interrater

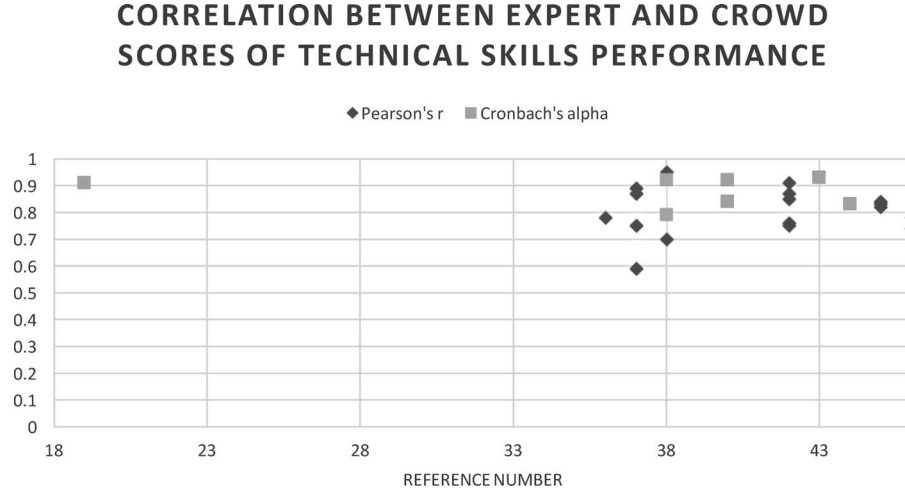

FIGURE 2
Summary of Crowd and Expert Correlations of Procedural Performance Across Included Studies

TABLE 1
Summary of Current Studies Evaluating Application of Crowd-Based Evaluation of Surgical Skills

Source, y	Task Performers	Total/Screened Evaluations[a]	Task
Chen et al,[39] 2014	1 *above average* performer	501/409 crowdworkers 107/67 Facebook users 10/9 attending surgeons	Robotic knot tying task
Holst et al,[19] 2015	3 urology residents (PGY-2, PGY-4, PGY-5) 2 urology faculty	250/206 crowd responses 50/34–43 crowdworkers/video 3 experienced robotic surgeons	FLS intracorporeal suturing
Holst et al,[43] 2015	12 surgeons of varying robotic surgical experience	600/487 crowd responses 50/35–46 crowdworkers/video 7 experienced robotic surgeons	Live, porcine, robotic-assisted urinary bladder closures
White et al,[40] 2015	49 surgeons ▪ 25 urology, general surgery, and obstetrics and gynecology trainees (PGY-1 to PGY-6) ▪ 24 faculty surgeons	2027/1443 crowd responses for pegboard task 30 crowdworkers/video 1668/1498 crowdworkers for suturing task 30 crowdworkers/video 3 experienced urologic surgeons	▪ Robotic rocking pegboard ▪ Robotic suturing task
Malpani et al,[41] 2015	4 expert surgeons 14 trainee surgeons	147 crowd responses 8 expert faculty surgeons 1 expert surgeon (assign global rating score)	Robotic suture throw and knot tying (evaluated by task segments)
Aghdasi et al,[44] 2015	26 participants in otolaryngology ▪ Medical students ▪ Residents ▪ Attending physicians	780 crowd responses 30 crowdworkers/video 3 expert faculty	Simulated cricothyroidotomy procedure
Polin et al,[42] 2016	105 participants in obstetrics and gynecology, urology, general surgery ▪ Trainees ▪ Fellows ▪ Surgeons	448 crowd responses 41 to 43 crowdworkers/video 3 expert robotic surgeons	Robotic surgical drills ▪ Tower transfer ▪ Roller coaster ▪ Big dipper ▪ Train tracks ▪ Figure-of-8
Vernez et al,[47] 2017	25 medical student urology residency interviewees	Open square knot tying: 1606 crowd responses/50 videos Laparoscopic peg transfer: 749 crowd responses Robotic suturing: 767 crowd responses; Skill task 8 on LAP mentor: 816 crowd responses 6 expert surgeon response; 2 experts/video	▪ Open square knot tying ▪ Laparoscopic peg transfer ▪ Robotic suturing ▪ Skill task 8 on LAP mentor
Deal et al,[36] 2016	7 general surgery intern volunteers	203 crowdworkers 6 faculty experts	FLS tasks ▪ Peg transfer ▪ Precision cutting ▪ Intracorporeal knot tying
Lee et al,[37] 2017	99 Canadian trainees ▪ Medical student urology applicants ▪ Urology trainees (PGY-3 and PGY-5) ▪ 6 attending urologists	No. of crowdworkers not reported 2 expert faculty	AUA basic laparoscopic urologic skills curriculum tasks ▪ Peg transfer ▪ Pattern cutting ▪ Suturing/knot tying ▪ Vascular clip application

TABLE 1
Summary of Current Studies Evaluating Application of Crowd-Based Evaluation of Surgical Skills (continued)

Source, y	Task Performers	Total/Screened Evaluations[a]	Task
Kowalewski et al,[38] 2016	24 representative videos of medical students, residents, fellows, and faculty from 8 academic urology training programs	1840/1438 crowd responses 60 crowdworkers per video 5 expert faculty	AUA basic laparoscopic urologic skills curriculum tasks ▪ Peg transfer ▪ Suturing
Powers et al,[45] 2016	5 surgeons ▪ PGY-3 and PGY-4 urology residents ▪ Attending surgeons	548 crowd responses $\geq$ 30 crowdworkers/video $\geq$ 3 clinical experts/video	Intraoperative renal artery and vein dissection during live robotic partial nephrectomy
Ghani et al,[46] 2016	Practicing urologists enrolled in MUSIC	30–55 crowd responses/video 25 MUSIC surgeons 4 peer reviewers/video	Live video from nerve-sparing robotic-assisted laparoscopic radical prostatectomy ▪ Bladder neck dissection ▪ Apical dissection ▪ Nerve sparing ▪ Urethrovesical anastomosis

Abbreviations: PGY, postgraduate year; FLS, fundamentals of laparoscopic surgery; AUA, American Urological Association; MUSIC, Michigan Urological Surgery Improvement Collaborative.

[a] Across all studies, all evaluators (total) were screened prior to inclusion. Those who did not meet validation standards were excluded from study analysis, and the remainder were included (screened).

reliability among expert scores was moderate to good overall, with a wide range of correlations: Cronbach's alpha ranged from 0.79 to 0.95, Krippendorff's alpha ranged from 0.25 to 0.55, intraclass correlation ranged from 0.38 to 0.88, and Fleiss' kappa = 0.55 (provided as online supplemental material).

Across nearly all studies, it was quicker to receive feedback from crowds than it was from experts. One article did not report time to feedback, and 2 reported that value for crowdworkers only.[37,40,42] For the remaining studies, average time to return feedback from nonexperts ranged from 2 hours and 50 minutes to 5 days, depending on the video length and task complexity. Most nonexperts responded within 48 hours. In contrast, experts took between 26 hours and 60 days to return feedback for the same tasks. Where reported, crowds consistently completed evaluations more quickly, ranging 4.8 to 150.9 times faster than experts (TABLE 2). In 1 study, it was noted that crowd responses were faster when the remuneration was doubled, suggesting that compensation may directly affect crowdsourcing efficiency.[40]

Where reported, remuneration for nonexpert evaluations was minimal, ranging from $0.25 to $1.00 per task (TABLE 2). One study used community volunteers rather than Amazon Mechanical Turk workers, which compensated participants with a $10 gift card.[41] Five studies did not report the crowdworker remuneration.[36,37,42,45,46] One study computed the cost difference for crowdworkers and experts to evaluate robotic pegboard transfer and suturing tasks.[40] Crowdworkers' costs were estimated at $16.50 for 30 evaluations versus $54 to $108 for 3 surgeon evaluations, suggesting that crowd-based feedback may be a more economical way to evaluate technical performance.

Discussion

Across multiple studies, there was considerable concordance between objective evaluation scores from crowds and experts for almost all tasks and task components examined. Moreover, crowd-based feedback was consistently more timely and less expensive than feedback from expert surgeons. These findings are consistent with those reported in a 2016 systematic review by Katz.[34] Our review additionally included studies that specifically examined novice as well as expert surgeons, representing the extremes of technical ability.[46,47]

As no previous studies have reported subanalyses for differing training levels, the inclusion of studies spanning a wider range of technical abilities allowed an initial assessment of how crowdsourcing technology may apply to trainees at varying levels of experience. Notably, in studies evaluating practicing surgeons exclusively, excellent correlation persisted.[46] However, some of the poorest agreement between crowd and expert scores was found in the study that exclusively involved medical students, who are novices.[47] This finding is corroborated by Aghdasi et al,[44] who examined instances where crowd and expert scores of a cricothyroidotomy simulation

TABLE 2
Summary of Time and Cost of Crowdsourced Feedback

Source, y	Average Time to Feedback		Ratio of Expert: Crowd Time To Feedback	Nonexpert Compensation, $/Task
	Nonexperts	Experts		
Chen et al,[39] 2014	▪ Turk workers: 5 d ▪ Facebook users: 25 d	24 d	4.8	1.00/HIT
Holst et al,[19] 2015	2 h 50 min	26 h	9.2	0.50/HIT
Holst et al,[43] 2015	4 h 28 min	14 d	4.5	0.75/HIT
White et al,[40] 2015	▪ Suturing task: 8 h 52 min ▪ Pegboard task: 108 h 48 min	N/A		0.25/pegboard task 0.50/suturing task
Malpani et al,[41] 2015	< 72 h	~672 h	21.6	10 gift card/survey
Aghdasi et al,[44] 2015	10 h	60 d	144	0.50/HIT
Polin et al,[42] 2016	16 h	N/A	N/A	N/A
Vernez et al,[47] 2017	▪ Open knot tying: 3 h 4 min ▪ Laparoscopic peg transfer: 3 h 3 min ▪ Robotic-assisted suturing: 3 h 26 min ▪ LAP mentor: 3 h 27 min	22 d	150.9	0.44/HIT
Deal et al,[36] 2016	19 h	10 d	12.6	N/A
Kowaleswski et al,[38] 2016	48 h	10 d	5	0.67/task
Powers et al,[45] 2016	11 h 33 min	13 d	27	N/A
Ghani et al,[46] 2016	▪ GEARS: 21 h ▪ RACE: 38 h	15 d	▪ GEARS: 17.1 ▪ RACE: 9.5	N/A

Abbreviations: HIT, human intelligence task; N/A, not applicable; GEARS, Global Evaluative Assessment of Robotic Skills; RACE, Robotic Anastomosis and Competency Evaluation.

differed by at least 5 points; experts considered 3 of 4 of these subjects *average* or *beginner* level. These findings suggest that crowdsourcing may be more accurate when applied in more skilled or advanced level surgeons. However, definitive conclusions are precluded by the small number of studies within the literature. Further work examining the effect of the task performer's skill level on crowd and expert evaluations is warranted.

It is notable that the lowest performance score correlation was reported for a laparoscopic fulguration task performed by medical students (Cronbach's alpha = 0.32).[47] In that case, crowd scores were compared with an automatically generated ranking score based on kinematic metrics, such as tool path length, applied instrument forces, instrument collisions, and task time, rather than the global "eyes-on" assessment that might be provided by experts. In addition to the inexperience of the medical students, the difference in outcome metrics may partially explain the low correlation seen in that study.[47] When expert and crowd assessments for laparoscopic simulation tasks have been compared with other automated metrics for time, tool path length, jerk cost, speed, economy of motion, and errors, there was a consistent, but generally weaker, correlation to crowd scores than expert scores.[38] It may be that crowds evaluate technical skill via an overall *gestalt* assessment of performance rather than consideration in aggregate of the multiple, specific technical aspects inherent in the nuanced approach of experts. The determinants of discrepancies between crowd and expert evaluations are an area for further investigation.

Some of the disagreement between crowd and expert evaluations may be explained by the wide range in interrater agreement among experts, which was particularly poor in the 2 studies assessing live surgical video.[45,46] Although all studies assumed expert surgeons provided a standard *ground truth* for comparison—that is, a set of parameters based on real-world observation against which all other metrics were evaluated—experts themselves may disagree on objective ratings of the same performance. Such discrepancies are compounded by the small sample sizes in experts' ratings reported in the literature, ranging from 2 to 9 experts per video, reducing the validity of an average expert-rated score. In contrast, the same videos were evaluated by at least 30 nonexperts, with the exception of 1 study where crowdworker numbers were not reported.[37] Although there did not appear to be any consistent correlations

between number of expert raters and interrater agreement among studies in this review, the number of overall studies was too small to draw any meaningful conclusions regarding such trends. The relative paucity of expert feedback in the studies in this review may reflect the nature of current feedback to trainees; future studies should assess the effect of greater parity in numbers between experts and crowdworkers to more accurately establish a ground truth for comparison.

Perioperative feedback from "master surgeons" tends to be formative, driven largely by attention to principles of deliberate practice, vertical transfer, and careful deconstruction of complex tasks for trainees; throughout an operative case, such feedback regarding resident progress is "direct and ongoing."[48] One of the main limitations of crowdsourced assessment scores is that they are largely summative rather than formative. Few studies solicited critiques or descriptive feedback from crowds. However, when pooled comments were obtained and examined, crowds and experts were found to discuss similar themes regarding efficiency, tissue handling, depth perception, and bimanual dexterity with a high level of congruence.[36] In providing specific comments for individual components of a complex task, crowds might serve a similar function as experts in providing more formative feedback to help trainees master the individual steps of a more complex procedure. Future research should focus on the nature and role of crowd comments in helping trainees progress their technical skills.

There are several limitations to this review. We searched primarily on the PubMed database and excluded non-English language articles. This may have resulted in exclusion of some relevant studies. Review and analysis of articles were performed primarily by a single reviewer, creating an additional source of potential bias.

Given the relatively novel use of crowdsourcing for surgical skills evaluation, the few published studies identified with multiple queries, and the lack of additional unique studies identified through discussion with subject matter experts, we think this is the most comprehensive review of the literature on the subject to date.

Several areas of future research emerged. Future studies should include greater numbers of expert reviewers to minimize interrater variability, and efforts should be made to better define the efficacy of this technology across the range of technical abilities. Specific work should also further identify and investigate drivers of discrepant performance scores assigned by crowds and experts. Automated metrics may serve as a helpful benchmark in this process. The quality and nature of crowdsourced comments remains to be explored, and the role of potentially formative verbal feedback in technical skills development is unknown. New applications of crowdsourced feedback remain to be defined, including areas with early use to date, such as early identification of surgically precocious trainees, selection of future residents, and peer coaching,[46,47] as well as the assessment of nontechnical skills, such as professionalism and communication. Crowdsourcing could also become an integral part of continuing medical education and skills development for surgeons in practice. With the adoption of milestone-based competencies for surgical training, such work will become particularly germane in the realm of surgical education.

Conclusion

Crowdsourcing in surgical education is a relatively novel phenomenon that may help address challenges in providing feedback in the current paradigm of surgical skills acquisition. The consistency, economics, and rapidity of crowd-based feedback make it readily scalable to large cohorts of trainees and surgeons. The validity of this technology across a wide breadth of procedural tasks and training levels makes it an appealing adjunct to the existing mechanisms of surgical skills feedback. Further work to better define an optimal population of trainees, elucidate limitations of crowdsourcing, and assess its use in the development of surgical skills over time are needed before crowdsourcing becomes effectively integrated within current surgical education paradigms.

References

1. Reznick RK, MacRae H. Teaching surgical skills—changes in the wind. *N Engl J Med*. 2006;355(25):2664–2669.
2. Polavarapu HV, Kulaylat AN, Sun S HO. 100 years of surgical education: the past, present, and future." *Bull Am Coll Surg*. 2013;98(7):22–27.
3. Institute of Medicine Committee on Quality of Health Care in America. Crossing the quality chasm: a new health system for the 21st century. In: Kohn LT, Corrigan JM, Donaldson MS, eds. *To Err is Human: Building a Safer Health System*. Washington, DC: National Academies Press; 2000.
4. Altman DE, Clancy C, Blendon RJ. Improving patient safety—five years after the IOM report. *N Engl J Med*. 2004;351(20):2041–2043.
5. Rogers Jr SO, Gawande AA, Kwaan M, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. *Surgery*. 2006;140(1):25–33.

6. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. 2013;369(15):1434–1442.
7. Hogg M, Zenati M, Novak S, et al. Grading of surgeon technical performance predicts postoperative pancreatic fistula for pancreaticoduodenectomy independent of patient-related variables. *Ann Surg*. 2016;264(3):482–491.
8. Boyle E, Al-Akash M, Gallagher AG, et al. Optimising surgical training: use of feedback to reduce errors during a simulated surgical procedure. *Postgrad Med J*. 2011;87(1030):524–528.
9. Bosse HM, Mohr J, Buss B, et al. The benefit of repetitive skills training and frequency of expert feedback in the early acquisition of procedural skills. *BMC Med Educ*. 2015;15(1):22.
10. Perone J, Fankhauser G, Adhikari D, et al. Who did the case? perceptions on resident operative participation. *Am J Surg*. 2017;213(4):821–826.
11. Morgan R, Kauffman DF, Doherty G, et al. Resident and attending perceptions of resident involvement: an analysis of ACGME reporting guidelines. *J Surg Educ*. 2017;74(3):415–422.
12. Reznick RK. Teaching and testing technical skills. *Am J Surg*. 1993;165(3):358–361.
13. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med*. 2003;15(4):270–292.
14. Gundle KR, Mickelson DT, Hanel DP. Reflections in a time of transition: orthopaedic faculty and resident understanding of accreditation schemes and opinions on surgical skills feedback. *Med Educ Online*. 2016;21(1):30584.
15. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273–278.
16. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190(1):107–113.
17. Goh AC, Goldfarb DW, Sander JC, et al. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. *J Urol*. 2012;187(1):247–252.
18. Shah J, Darzi A. Surgical skills assessment: an ongoing debate. *BJU Int*. 2001;88(7):655–660.
19. 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(10):150413093359007.
20. Stefanidis D, Arora S, Parrack DM, et al. Research priorities in surgical simulation for the 21st century. *Am J Surg*. 2012;203(1):49–53.
21. Howe J. The rise of crowdsourcing. *Wired Mag*. 2006;14(6):1–5.
22. Garrigos-Simon FJ, Gil-Pechuán I, Estelles-Miguel S, eds. *Advances in Crowdsourcing*. Cham, Switzerland: Springer International Publishing AG; 2015.
23. Brabham DC. Crowdsourcing as a model for problem solving: an introduction and cases. *Convergence*. 2008;14(1):75–90.
24. Cooper S, Khatib F, Treuille A, et al. Predicting protein structures with a multiplayer online game. *Nature*. 2010;466:756–760.
25. Kawrykow A, Roumanis G, Kam A, et al. Phylo: a citizen science approach for improving multiple sequence alignment. *PLoS One*. 2012;7(3):e31362.
26. Nguyen TB, Wang S, Anugu V, et al. Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. *Radiology*. 2012;262(3):824–833.
27. Mavandadi S, Dimitrov S, Feng S, et al. Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study. *PLoS One*. 2012;7(5):e37245.
28. Chen SP, Kirsch S, Zlatev D V, et al. Optical biopsy of bladder cancer using crowd-sourced assessment. *JAMA Surg*. 2016;151(1):90–92.
29. Brady CJ, Villanti AC, Pearson JL, et al. Rapid grading of fundus photographs for diabetic retinopathy using crowdsourcing. *J Med Internet Res*. 2014;16(10):e233.
30. Bow HC, Dattilo JR, Jonas AM, et al. A crowdsourcing model for creating preclinical medical education study tools. *Acad Med*. 2013;88(6):766–770.
31. Blackwell KA, Travis MJ, Arbuckle MR, et al. Crowdsourcing medical education. *Med Educ*. 2016;50(5):576.
32. Reed DA, Cook DA, Beckman TJ, et al. Association between funding and quality of published medical education research. *JAMA*. 2007;298(9):1002.
33. Lendvay TS, White L, Kowalewski T. Crowdsourcing to assess surgical skill. *JAMA Surg*. 2015;150(11):1086–1087.
34. Katz AJ. The role of crowdsourcing in assessing surgical skills. *Surg Laparosc Endosc Percutan Tech*. 2016;26(4):271–277.
35. White LW, Lendvay TS, Holst D, et al. Using crowd-assessment to support surgical training in the developing world. *J Am Coll Surg*. 2014;219(4, suppl):e40.
36. Deal SB, Lendvay TS, Haque MI, et al. Crowd-sourced assessment of technical skills: an opportunity for improvement in the assessment of laparoscopic surgical skills. *Am J Surg*. 2016;211(2):398–404.
37. Lee JY, Andonian S, Pace KT, et al. Basic laparoscopic skills assessment study—validation and standard setting among Canadian urology trainees. *J Urol*. 2017;197(6):1539–1544.
38. Kowalewski TM, Comstock B, Sweet R, et al. Crowd-sourced assessment of technical skills for validation of

basic laparoscopic urologic skills tasks. *J Urol*. 2016;195(6):1859–1865.
39. 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(1):65–71.
40. White LW, Kowalewski TM, Dockter RL, et al. Crowd-sourced assessment of technical skill: a valid method for discriminating basic robotic surgery skills. *J Endourol*. 2015;29(11):1295–1301.
41. Malpani A, Vedula SS, Chen CCG, et al. A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. *Int J Comput Assist Radiol Surg*. 2015;10(9):1435–1447.
42. Polin MR, Siddiqui NY, Comstock BA, et al. Crowdsourcing: a valid alternative to expert evaluation of robotic surgery skills. *Am J Obstet Gynecol*. 2016;215(5):644.e1–644.e7.
43. 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(5):604–609.
44. Aghdasi N, Bly R, White LW, et al. Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. *J Surg Res*. 2015;196(2):302–306.
45. Powers MK, Boonjindasup A, Pinsky M, et al. Crowdsourcing assessment of surgeon dissection of renal artery and vein during robotic partial nephrectomy: a novel approach for quantitative assessment of surgical performance. *J Endourol*. 2016;30(4):447–452.
46. Ghani KR, Miller DC, Linsell S, et al. Measuring to improve: peer and crowd-sourced assessments of technical skill with robot-assisted radical prostatectomy. *Eur Urol*. 2016;69(4):547–550.
47. Vernez SL, Huynh V, Osann K, et al. C-SATS: assessing surgical skills among urology residency applicants. *J Endourol*. 2017;31(suppl 1):95–100.
48. McKendy KM, Watanabe Y, Lee L, et al. Perioperative feedback in surgical training: a systematic review. *Am J Surg*. 2016;214(1):117–126.

Jessica C. Dai, MD, is a Urology Resident, Department of Urology, University of Washington; **Thomas S. Lendvay, MD,** is Pediatric Urology Fellowship Program Director, Seattle Children's Hospital; and **Mathew D. Sorensen, MD, MS,** is Urology Residency Program Director, Department of Urology, University of Washington.

Funding: This work was supported by the VA Puget Sound Health Care System, Seattle, Washington.

Conflict of interest: Dr. Lendvay is co-founder and chief medical officer of C-SATS Inc, a commercially available platform for surgical skills improvement based on the technology described within this review.

Corresponding author: Jessica C. Dai, MD, University of Washington, Department of Urology, HSB BB 1121, 1959 NE Pacific Street, Seattle, WA 98195, 206.685.1982, jcdai@uw.edu

Received May 1, 2017; revision received July 26, 2017; accepted August 7, 2017.