

Modeling Performance of Open Surgical Cases

David P. Azari, PhD;

Lane L. Frasier, MD, MS;

Brady L. Miller, MD, MPH;

Sudha R. Pavuluri Quamme, MD, MS;

Brian V. Le, MD, MA;

Caprice C. Greenberg, MD, MPH,
FACS;

Robert G. Radwin, PhD

Introduction: Previous efforts used digital video to develop computer-generated assessments of surgical hand motion economy and fluidity of motion. This study tests how well previously trained assessment models match expert ratings of suturing and tying video clips recorded in a new operating room (OR) setting

Methods: Enabled through computer vision of the hands, this study tests the applicability of assessments born out of benchtop simulations to in vivo suturing and tying tasks recorded in the OR.

Results: Compared with expert ratings, computer-generated assessments for fluidity of motion (slope = 0.83, intercept = 1.77, $R^2 = 0.55$) performed better than motion economy (slope = 0.73, intercept = 2.04, $R^2 = 0.49$), although 85% of ratings for both models were within ± 2 of the expert response. Neither assessment performed as well in the OR as they did on the training data. Assessments were sensitive to changing hand postures, dropped ligatures, and poor tissue contact—features typically missing from training data. Computer-generated assessment of OR tasks was contingent on a clear, consistent view of both surgeon's hands.

Conclusions: Computer-generated assessment may help provide formative feedback during deliberate practice, albeit with greater variability in the OR compared with benchtop simulations. Future work will benefit from expanded available bimanual video records.

(*Sim Healthcare* 16:e188–e193, 2021)

Key Words: Automated surgical skill assessment, computer vision, surgical performance models.

The purpose of this study is to test the accuracy of computer-generated assessments of surgical dexterity in the operating room (OR) during common suturing and tying tasks. The computer-generated assessments are previously trained linear and generalized additive models¹ that ingest a kinematic record of hand motion and produce measures along 2 previously tested² scales (Fig. 1) of motion economy (ie, efficiency, 0–10) and fluidity of motion (ie, smoothness, 0–10). Each model is trained on expert ratings of benchtop suturing and tying. This study tests how well the assessments match an expert panel ($n = 3$) consensus score for suturing and tying in the OR. This article provides a discussion on the accuracy and utility of these tools, as well as implications for future training programs in expanding computer-generated surgical assessment.

From the Department of Industrial and Systems Engineering (D.P.A., R.G.R.); Department of Surgery (S.R.P.Q., C.C.G.), Clinical Sciences Center; Department of Urology (B.V.L.); and Duane H. and Dorothy M. Bluemke Professor in the College of Engineering (R.G.R.), University of Wisconsin-Madison, Madison, WI; Department of Surgery (L.L.F.), Penn Medicine - University of Pennsylvania Health System, Philadelphia, PA; City of Hope National Comprehensive Cancer Center (B.L.M), Duarte, CA.

Correspondence to: Robert G. Radwin, PhD, Department of Industrial and Systems Engineering, University of Wisconsin-Madison, 1550 Engineering Dr, Madison, WI 53706 (e-mail: rradwin@wisc.edu).

The authors declare no conflict of interest.

Supported by the Society of Academic Urologists, Grant number MSN205847.

Previous presentations: Azari DP, Miller BL, Le BV., et al. Can surgical performance for varying experience be measured from hand motions? *Proc Hum Factors Ergon Soc Annu Meet.* 2018;62(1):583–587.

All actions in this study were approved by the University of Wisconsin-Madison Education and Social/Behavioral Sciences Institutional Review Board.

Complementary work develops additional models that rely on similar data, but for different purposes.^{1,2,33,38,41}

Copyright © 2021 Society for Simulation in Healthcare

DOI: 10.1097/SIH.0000000000000544

Assessing Surgery

Surgical performance consists of various domains³ that are difficult to assess^{4–7} and characterize.⁸ The most widely accepted measure of surgical psychomotor performance—the Objective Structured Assessment of Technical Skills (OSATS)⁹—consists of a series of global rating scales and procedural checklists for 8 prescribed simulations. Studies of OSATS have consistently validated its use for formative feedback in simulated settings.¹⁰ Robotic platforms, like Robotic Video and Motion Analysis Software (ROVIMAS),¹¹ and the Imperial College Surgical Assessment Device (ICSAD),^{12–14} and virtual reality simulations,¹⁵ among others, provide ready fodder for using OSATS and objective computer-aided technical skill evaluation¹⁶ research as part of a growing body of “surgical data science.”¹⁷ The ICSAD, in particular, has shown positive correlation among hand motion and global rating scales, significantly discriminating between experienced and novice performers.¹⁸ Complimentary efforts to measure performance on robotic video data are increasingly leveraging deep learning methods like linear classifiers, support vector machines, and random forests¹⁹ as well as neural networks²⁰ and convolutional neural networks.²¹ Such efforts are assisted by growth and availability in broad and accessible data sets like the Johns Hopkins University–Intuitive Surgical Gesture and Skill Assessment Working Set.²² The Dextrous Surgical Skills Transfer Dataset²³—made available in 2019—focuses on cross-platform comparisons.

Despite these advances in robotic surgical data science, objective computer-aided technical skill evaluation has not been consistently applied to bimanual observation and assessment in the OR.¹⁶ This represents a gap in extrapolating assessment scores from simulated settings to dynamic real-world conditions—a crucial component of modern approaches to

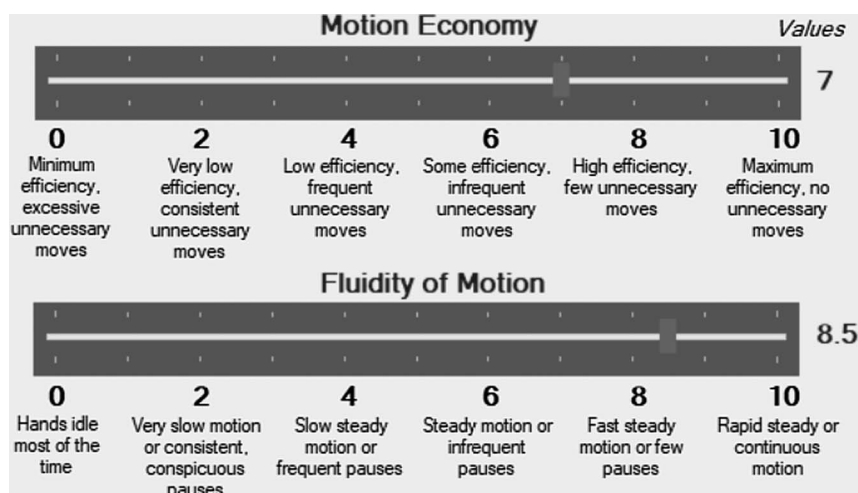


FIGURE 1. Expert rating scales of motion economy (ie, efficiency, 0–10) and fluidity of motion (ie, smoothness, 0–10).

validity advocated by Michael Kane^{24,25} and further explicated by David Cook^{26,27} to guide development of surgical assessments.

Marker-less video tracking has demonstrated the ability to observe differences in hand motion between clinicians in and out of the OR^{28–30} and to predict real-world performance along a series of subjective rating scales.² There is continuing interest in applying sensor-free computer vision³¹ and exploring how merging sensor-equipped hand movement records with computer vision³² can improve various predictions. As a result, we are in a unique position to leverage marker-less motion tracking of hand movements to automate OSATS-like assessments—in different settings, and of clinicians of varying experience—to begin assembling a basis of validity evidence in line with Kane's approach.²⁴

This study aims to extend previous work developing computer-generated assessments of clinician hand motion during benchtop simulations to the OR—exploring how well assessments of “performances in real life”¹⁰ match those from more-controlled settings.

METHODS

Participants and Settings

This study relies on video data collected both from benchtop simulations (Fig. 2A) and previously recorded in OR (Fig. 2B) by Azari et al.² Expert ratings of motion economy and fluidity of motion during benchtop simulations served as training data for the computer-generated assessments, whereas expert ratings of OR video now serves as testing data. Participation and

recruitment in both settings were approved by the University of Wisconsin-Madison Education and Social/Behavioral Sciences Institutional Review Board. Residents who completed less than 4 postgraduate years and surgeons with less than 7 years as attendings were classified as “junior” within their position (Table 1).

For the benchtop setting, surgeons and students ($n = 37$) from the University of Wisconsin Hospital and Clinics were recruited to participate through e-mail list-servs, in-person announcements, and personal recommendation. Each clinician performed a series of simple interrupted sutures and a running subcuticular suture on an Allevyn Hydrocellular foam dressing (produced by Smith+Nephew, <http://www.smith-nephew.com>) mounted to a wood block for stability (Fig. 2A). Before participation, each clinician completed a basic demographic survey specifying their years of experience, handedness (ie, right- or left-handed), specialty, and amount of training. Each video clip was calibrated to real-world distances using a notecard, ruler, or suture packet of known size. A total of 2 hours 42 minutes of active suturing and tying segments yielded 219 training clips (Table 1). This collection and video set are further described by Azari et al.³³

Video collection of surgeons in the OR ($n = 7$) was enabled by remotely operated video recorder mounted to the overhead lamp (Fig. 3A). In this setting, the proximal interphalangeal joint breadths for male and female patients (Fig. 3B) provided calibration when no object of known size was visible. We have previously used hand³⁵ and joint²⁸ breadth measurements with

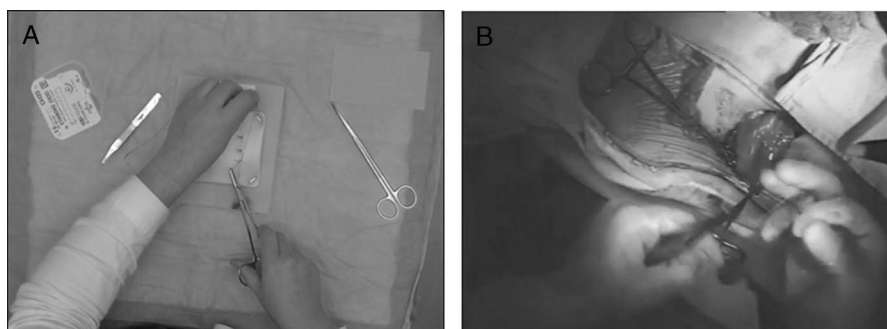


FIGURE 2. Top-down view of common suturing tasks on foam (A) and in the OR (B).

TABLE 1. Number of Participants and Length of Video Recorded by Task and Material

Material	Task	MS	JR	SR	JA	SA	RT	Total	Video (hh:mm:ss)
Foam (A)	SI, RS	10	5	10	4	6	2	37	02:42:00
OR (B)	S, T	0	0	2	2	3	0	7	00:10:19
Totals (A + B)		10	5	12	6	9	2	44	02:52:19

JA, junior attending; JR, junior resident; MS, medical student; RS, running subcuticular suturing; RT, retired; S, general suturing; SA, senior attending; SI, simple interrupted suturing; SR, senior resident; T, general tying.

low coefficients of variation³⁴ as calibration coefficients. Pixel locations were identified through Multimedia Video Task Analysis software specially developed at the University of Wisconsin-Madison,³⁶ allowing us to calibrate pixel measurements in the video to real-world units. A total of 10 minutes 19 seconds of footage from all participants yielded 48 clips of bimanual active suturing and tying tasks. Expert ratings on these tasks comprised the test data for each computer-generated assessment.

Motion Tracking

We used custom software³⁷ to track a region of interest in each video frame over the course of a video (Fig. 4). We have previously used this technology to observe differences between dominant and nondominant hands,²⁹ identify differences between tasks and roles,²⁸ summarize observed kinematic differences between clinicians of varying tenure,³⁸ and predict expert-rated performance of the dominant hand during short video clips in the OR.² This study extends the previous exploratory modeling techniques to test linear and generalized additive models trained using controlled benchtop settings with a diverse participant group into the more dynamic setting of the OR to include shifting postures, intermittent views, and tissue variability.

Expert Ratings

For the OR videos, a panel of expert surgeons ($n = 3$) rated clips individually, discussed their ratings, and arrived at a consensus score.² These consensus scores were then used as benchmark calibrations for later benchtop ratings. For the benchtop setting, a panel of expert surgeons ($n = 5$) rated 219 clips (2 hours 42 minutes) of active suturing or tying on foam via a custom USB rating applet. The mean of ratings was maintained in the training set if all ratings fell within a range of 3. If the range of ratings among the panel was greater than 3, either (1) 3 of 4 ratings needed to fall within a range of 2 or (2) 4 of 5 ratings fell within a range of 2.5. The rating procedure, collection scheme, and approach to outlier removal are previously published by Azari et al.^{2,33} Regardless of setting, each rating surgeon had at least 3 years of experience in the attending role and was involved in educational and coaching efforts.

Predictive Models

Computer-generated assessments were trained on expert ratings from the benchtop setting. A GAM was selected to predict the consensus (mean) of panel ratings for motion economy, whereas a linear model was selected to predict the consensus (mean) of panel ratings for fluidity of motion. In contrast to linear least-square regression models, generalized additive models^{39,40} rely on summing a series of smoothed (nonparametric) functions that better incorporate nonlinear effects into the overall prediction. The features of hand motion (eg, speed,

acceleration, etc) served as independent variables and were selected using common penalized regression shrinkage methods (eg, Ridge, LASSO, and Elastic Net) and balanced by examining the mean-square error and Akaike information criterion. Azari¹ further details the feature selection, refinement, and internal training of these models.

Assessment Accuracy

The accuracy of the computer-generated assessments in the OR is compared against their previous accuracy in the benchtop setting. An ideal result would be similar intercepts (close to 0), slopes (close to 1), and high R^2 (>0.7) values when ground truth (ie, expert rated) and predicted (ie, computer-generated) assessments are plotted on orthogonal axes.

RESULTS

A total of 48 OR clips of common suturing and tying tasks (mean length = 13.06 seconds) met inclusion criteria to test both computer-generated assessments. These videos exhibited both: (1) a clear view of both hands for the entire clip and (2) observable movement among both hands.

When tested on the OR videos, the computer-generated assessments of motion economy and fluidity of motion derived from Azari¹ predicted the expert ratings moderately well, with mean squared errors of 1.74 and 1.58, respectively. For computer-generated assessment of motion economy (Fig. 5, slope = 0.73, intercept = 2.04, $R^2 = 0.49$), 85% of residuals fell within ± 2 of the linear relationship, while for fluidity of motion (Fig. 6, slope = 0.83, intercept = 1.75, $R^2 = 0.55$), 86% of residuals fell within ± 2 of the linear relationship.

Notable outliers in the motion economy assessments (Fig. 5) at $x = 2$, $y = 5.9$ (A), $x = 4$, $y = 2.6$ (B), and $x = 4$, $y = 1.2$ (C) exhibited the following characteristics: (1) multiple missed suturing attempts while driving a needle, (2) a dropped ligature, and (3) changing position and posture to access tissue.

Inspection of Figure 6 highlights one substantial fluidity of motion underprediction at $x = 5$, $y = 2.5$. In this outlier clip, the surgeon changes posture twice before beginning the suture, first by reaching across the patient, and then returning



FIGURE 3. Operating room light-mounted camera system (A) and pixel to real-world calibration (B) using proximal interphalangeal joint breadth (PIJB)³⁴ if no standard markers are visible.

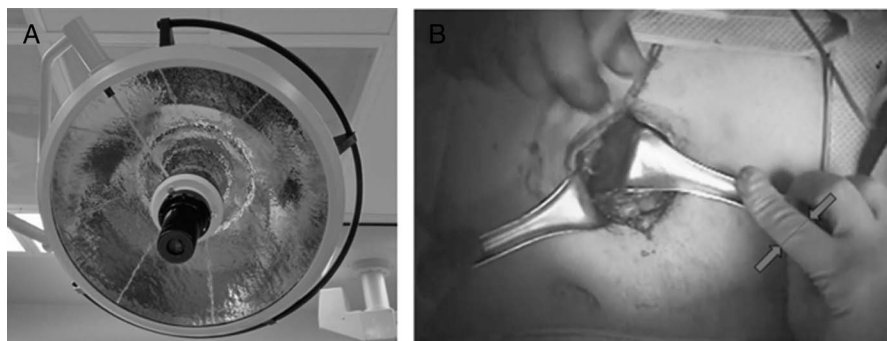


FIGURE 4. Region of interest to track motion of participant's nondominant hand while operating.

to the original position, ostensibly to gain the best angle of access to ensure appropriate tension on friable bowel tissue.

Neither computer-generated assessment was as accurate in the OR as they were on the benchtop simulations. As described in Azari,¹ both the motion economy assessment (slope = 0.65, intercept = 2.36, $R^2 = 0.66$) and fluidity of motion assessment (slope = 0.71, intercept = 1.98, $R^2 = 0.77$) exhibited reduced R^2 values when applied in the OR (0.49, 0.55, respectively), despite similar intercepts and slopes compared with the training data. The computer-generated assessment of fluidity of motion more closely mirrored the OR expert ratings than its motion economy counterpart.

DISCUSSION

This study tested how well previously trained computer-generated assessments extrapolate to suturing and tying tasks in the OR. Both benchtop-derived assessments (ie, motion economy, fluidity of motion) exhibited lower R^2 values on the real-world data, suggesting that the benchtop setting training data may not account for contextual variation (eg, changing postures, positions, errors, etc) in the OR. Outlier investigation showed instances like missing the tissue while driving a needle, moving to better fit the hands inside the body cavity, and dropping or resetting ligatures on slippery, friable tissue—features not available in the training data. These variations indicate a deviation in hand features that otherwise inform the computer-generated assessment and suggest further utility in expanding the available bimanual video training data.

Outlier examination also confirms that the computer-generated assessments make no distinction between errors and intentional movements. For instance, the automatic predictions do not account for contextual features of the surgery that drive changes in hand path. A participant suturing on foam, for instance, may drive a needle too deeply into the foam but continue to complete the exercise, whereas an experienced surgeon in the OR may prefer to reset the needle driver instead of completing a risky suture. The motion tracking algorithm makes no distinction between these or other maneuver techniques, such as one- or two-handed or tool-based tying—complicating the feature set that feeds the assessment. These contextual features have significant implications for live surgery and minimal implications in training settings. Even so, training gains must manifest appropriately within the subsequent live context, requiring the surgeon in training to adapt and apply in potentially new ways the techniques and movements learned deliberately in while practicing.

This study was limited to suturing and tying tasks and did not comprise a holistic assessment of a surgeon's experience or performance in all domains³ of surgical skill. An observed delay in hand movement while suturing on foam could indicate a lack of automaticity or underdeveloped mental schema to sufficiently represent the task—prompting clear training remediation. In the OR, however, where access to the body is constrained, deliberate or purposeful postural changes may be indicative of greater cross-domain synthesis and familiarity with the task constraints, in which the surgeon delays to accommodate sensitive tissue. An expert rater may be inclined to recognize this context and unintentionally afford the participant a heightened fluidity score, thereby increasing the variability in the testing data. On the other hand, repeated changes in posture or balance in the OR could indicate a lack of confidence or preparedness, although those motions could be considered efficient. Future advances in OR video recording may offset challenges in connecting surgical skills domains with psychomotor behavior and the observable hand movements by building out a more well-rounded bank of surgical video from which to test and train computer-generated assessments.

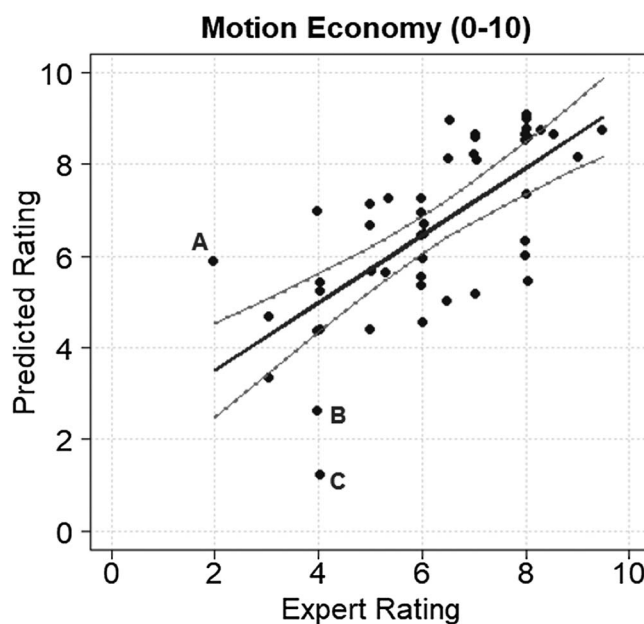


FIGURE 5. Predicted ratings versus expert ratings for motion economy (0–10) rating scale using ($n = 48$) video clips from the OR. Confidence intervals (95%) are shown on either side of the linear fit.

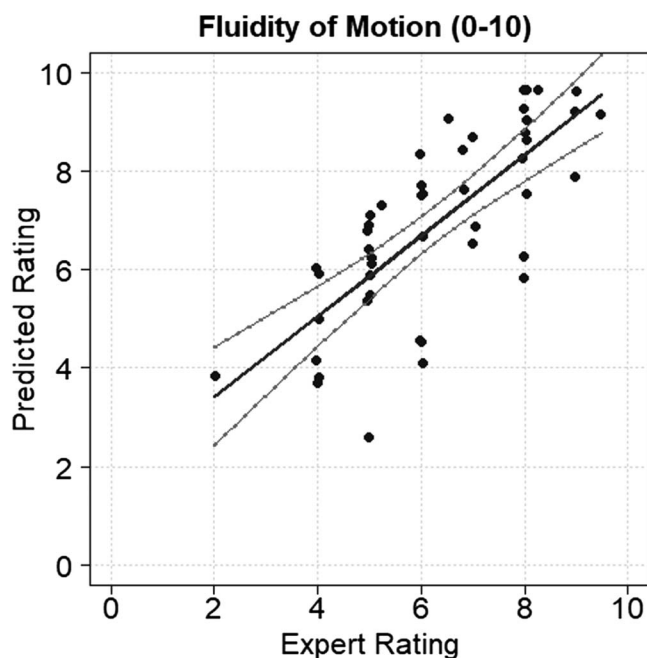


FIGURE 6. Predicted ratings versus expert ratings for fluidity of motion (0–10) rating scale using ($n = 48$) video clips from the OR. Confidence intervals (95%) are shown on either side of the linear fit.

In a similar vein, this study does not differentiate between tissue types or suturing and tying maneuvers. Future expansion of digital video records and data mining may alleviate this limitation. Parallel work⁴¹ has automatically classified suturing and tying periods every 2 seconds with 79% accuracy and may be used in the future to further refine measures of fluidity of motion and motion economy within the simulated setting. Future work may also explore how different simulated tissues (eg, foam, balloons, tissue paper) affect assessment scores. Segmenting video by tissue type, as we have explored for single-handed assessment routines in the past,²⁸ may improve overall accuracy by accommodating different behaviors for different tissues. However, this increases the burden on manual labeling of video data, decreases power, and limits the potential for comparison between short clips observed in the OR and video records collected during benchtop trials.

The different methods in deriving a consensus score for OR videos (ie, consensus approach²) and benchtop videos (ie, calibration, outlier removal¹) reduce consistency between the expert-generated and predicted ratings across the training and testing data. This challenge was mitigated by using the initial consensus ratings and video examples to calibrate all subsequent raters.

Evaluating assessment accuracy was also limited by the experience required to be a practicing clinician and the available number of video clips ($n = 48$) derived from a limited number of surgeons operating in the OR ($n = 7$). Although video training clips ($n = 219$) included records from students, residents, and attendings suturing and tying on foam, only senior residents and attendings may be approved to operate. The automatic routines tested here, however, may provide as a formative assessment resource for more junior clinicians, reducing the burden on experienced surgeons to provide feedback and better preparing

medical students and residents for evaluation. Despite these limitations, outlier investigation suggests clear contributors of prediction variability, including tissue type, shifting or constrained postures, and contextual variation natural to surgical cases.

Future model and video examination may be able to “reverse engineer” the specific features of movement, which are contributing to overprediction or underprediction at any point in a surgical record. A “reliability index” in the prediction may be able to limit outliers by presenting each prediction with a confidence estimate. These processing additions may be most applicable during formative exercises, to express the relative confidence of the prediction model to the participant in training with minimal danger and barriers to implementation. Feature analysis could help also explore which motion properties are most salient and important to different raters and/or coaches. Coaches and trainers may benefit from working with students who have already practiced suturing and tying with the support of automated feedback about how their fluidity of motion and motion economy may be perceived in a real-world OR.

The results of this study emphasize 2 things: first, that the computer-generated assessment are moderately predictive of expert ratings in the OR and, second, that they are sensitive to context within that more complex environment, and produce outliers under changes in posture, errors, or multiple attempts. The range of residuals (± 2) of the computer-generated assessments was also greater than the original threshold for consensus imposed upon the expert panel at large. As a result, computer-generated assessment should be limited to providing formative feedback for training or retroactive investigation. The automatic routines could be packaged into a “digital assistant”—repeatedly used without sensors—and promote skills gain through deliberate practice without burdening experienced evaluators.

CONCLUSIONS

This study tests how well computer-generated assessments trained during benchtop simulations perform in the more variable and complex real-world OR. Expert video ratings serve as ground truth, and results are framed within the modern validity framework as articulated by Kane.²⁴ The results suggest that although computer vision of the hands and computer-generated assessments of common benchtop suturing tasks could provide automatic, quantitative feedback of medical student and resident performance, the same assessments do not account for the wider variation, contextual factors, or errors that expert surgeons recognize in the OR. Future work expanding available video and hand-motion data in various contexts will support continued refinement of automated assessment.

ACKNOWLEDGMENTS

The authors thank Trenton Feda, Akshat Khanna, Oguz Akkas, Renee Greene, Barbara Lewis, Connor Feldman, Carla Pugh, Jacob Greenberg, Reginald Bruskewitz, Kristin Long, and Amber Shada for their contributions to this work.

REFERENCES

1. Azari DP, Miller BL, Le B.V, et al. A comparison of expert ratings and marker-less hand tracking along OSATS-derived motion scales. *IEEE*

Transactions on Human-Machine Systems 2020;1–10. Available at: <https://doi.org/10.1109/THMS.2020.3035763>.

2. Azari DP, Frasier LL, Quamme SRP, et al. Modeling surgical technical skill using expert assessment for automated computer rating. *Ann Surg* 2019;269(3):574–581.
3. Madani A, Vassiliou MC, Watanabe Y, et al. What are the principles that guide behaviors in the operating room? *Ann Surg* 2017;265(2):255–267.
4. Moorthy K, Munz Y, Sarker SK, Darzi A. Objective assessment of technical skills in surgery. *Br Med J* 2003;327(7422):1032–1037.
5. Hopmans CJ, Den Hoed PT, Van Der Laan L, et al. Assessment of surgery residents' operative skills in the operating theater using a modified Objective Structured Assessment of Technical Skills (OSATS): a prospective multicenter study. *Surg (United States)* 2014;156(5):1078–1088.
6. Williams RG, Kim MJ, Dunnington GL. Practice guidelines for operative performance assessments. *Ann Surg* 2016;264(6):934–948.
7. Jelovsek JE, Kow N, Diwadkar GB. Tools for the direct observation and assessment of psychomotor skills in medical trainees: a systematic review. *Med Educ* 2013;47(7):650–673.
8. Azari DP, Greenberg CC, Pugh CM, Wiegmann DA, Radwin RG. In search of characterizing surgical skill. *J Surg Educ* 2019;76(5):1348–1363.
9. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg* 1997;173(3):226–230.
10. Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. *Adv Heal Sci Educ* 2015;20(5):1149–1175.
11. Aggarwal R, Grantcharov T, Moorthy K, et al. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. *Ann Surg* 2007;245(6):992–999.
12. Hayter MA, Friedman Z, Bould MD, et al. Validation of the Imperial College Surgical Assessment Device (ICSAD) for labour epidural placement. *Can J Anesth* 2009;56(6):419–426.
13. Datta V, Chang A, Mackay S, Darzi A. The relationship between motion analysis and surgical technical assessments. *Am J Surg* 2002;184(1):70–73.
14. Bann SD, Khan MS, Darzi AW. Measurement of surgical dexterity using motion analysis of simple bench tasks. *World J Surg* 2003;27(4):390–394.
15. Grober ED, Roberts M, Shin EJ, Mahdi M, Bacal V. Intraoperative assessment of technical skills on live patients using economy of hand motion: establishing learning curves of surgical competence. *Am J Surg* 2010;199(1):81–85.
16. Vedula SS, Ishii M, Hager GD. Objective assessment of surgical technical skill and competency in the operating room. *Annu Rev Biomed Eng* 2017;19(1):301–325.
17. Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering* 2017;1(9):691–696. Available at: <https://doi.org/10.1038/s41551-017-0132-7>.
18. Corvetto MA, Fuentes C, Araneda A, et al. Validation of the imperial college surgical assessment device for spinal anesthesia. *BMC Anesthesiol* 2017;17(1):1–7.
19. Lee D, Yu HW, Kwon H, Kong HJ, Lee KE, Kim HC. Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations. *J Clin Med* 2020;9(6):1964.
20. Khalid S, Goldenberg M, Grantcharov T, Taati B, Rudzicz F. Evaluation of deep learning models for identifying surgical actions and measuring performance. *JAMA Netw Open* 2020;3(3):e201664.
21. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. *Int J Comput Assist Radiol Surg* 2019;14(9):1611–1617.
22. Gao Y, Vedula SS, Reiley CE, et al. JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): a surgical activity dataset for human motion modeling. *Model Monit Comput Assist Interv – MICCAI Work* 2014;1–10.
23. Madapana N, Low T, Voyles RM, et al. DESK: a robotic activity dataset for dexterous surgical skills transfer to medical robots. *IEEE Int Conf Intell Robot Syst* 2019;Taurus II:6928–6934.
24. Kane M. The argument-based approach to validation. *School Psych Rev* 2013;42(4):448–457.
25. Kane M. Validation. In: Brennan R, ed. *Educational Measurement*. 4th ed. Westport, CT: American Council on Education and Praeger; 2006:17–64.
26. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Heal Sci Educ* 2014;19(2):233–250.
27. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ* 2015;49(6):560–575.
28. Frasier LL, Azari DP, Ma Y, et al. A marker-less technique for measuring kinematics in the operating room. *Surgery* 2016;160(5):1400–1413.
29. Glarner CE, Hu YY, Chen CH, et al. Quantifying technical skills during open operations using video-based motion analysis. *Surg (United States)* 2014;156(3):729–734.
30. Mackenzie CF, Watts D, Patel RR, et al. Sensor-free computer vision hand-motion entropy and video analysis of technical performance during open vascular surgery: proof of concept for methodology. *J Am Coll Surg* 2016;223(4):e63.
31. Mackenzie CF, Watts D, Patel R, et al. Sensor-free computer-vision hand-motion entropy and video-analysis of technical performance during open surgery on fresh cadavers: report of methodology and analysis. In: *Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting*. Washington DC; 2016:691–695.
32. Zia A, Sharma Y, Bettadapura V, Sarin EL, Essa I. Video and accelerometer-based motion analysis for automated surgical skills assessment. *Int J Comput Assist Radiol Surg* 2018;13(3):443–455.
33. Azari DP, Miller BL, Le BV, et al. Can surgical performance for varying experience be measured from hand motions? *Proc Hum Factors Ergon Soc Annu Meet* 2018;62(1):583–587.
34. Greiner TM. Hand anthropometry of U.S. army personnel. *Tech Rep Natick* 1991;TR-92/011:434.
35. Akkas O, Azari DP, Chen C-HE, et al. A hand speed – duty cycle equation for estimating the ACGIH hand activity level rating. *Ergonomics* 2015;58(2):184–194.
36. Yen TY, Radwin RG, M.F. S. Usability testing by multimedia video task analysis. In: Winters JM, ed. *Medical Instrumentation: Accessibility and Usability Considerations*. Boca Raton, FL: CRC Press; 2007:159–172.
37. Chen C-H, Hu YH, Radwin RG. A motion tracking system for hand activity assessment. In: *2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*. Xi'an, China, IEEE; 2014:320–324.
38. Azari DP, Miller BL, Le BV, Greenberg CC, Radwin RG. Quantifying surgeon maneuvers across experience levels through marker-less hand motion kinematics of simulated surgical tasks. *Appl Ergon* 2020;87:103136.
39. Chouldechova A, Hastie T. Generalized additive model selection. *Ann Appl Stat* 2015;1–24. Available at: <http://arxiv.org/abs/1506.03850>. Accessed date June 17, 2015.
40. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. *Math Intell* 2001;27(2):83–85.
41. Azari DP, Hu YH, Miller BL, Le BV, Radwin RG. Using surgeon hand motions to predict surgical maneuvers. *Hum Factors J Hum Factors Ergon Soc* 2019;61(8):1326–1339.