Objective: Computer vision was used to predict expert performance ratings from surgeon hand motions for tying and suturing tasks. 

Summary Background Data: Existing methods, including the objective structured assessment of technical skills (OSATS), have proven reliable, but do not readily discriminate at the task level. Computer vision may be used for evaluating distinct task performance throughout an operation.

Methods: Open surgeries were videoed and surgeon hands were tracked without using sensors or markers. An expert panel of 3 attending surgeons rated tying and suturing video clips on continuous scales from 0 to 10 along 3 task measures adapted from the broader OSATS: motion economy, fluidity of motion, and tissue handling. Empirical models were developed to predict the expert consensus ratings based on the hand kinematic data records.

Results: The predicted versus panel ratings for suturing had slopes from 0.73 to 1, and intercepts from 0.36 to 1.54. Predicted versus panel ratings for tying had slopes from 0.39 to 0.88, and intercepts from 0.79 to 4.36. The mean square error among predicted and expert ratings was consistently less than the mean squared difference among individual expert ratings and the eventual consensus ratings.

Conclusions: The computer algorithm consistently predicted the panel ratings of individual tasks, and were more objective and reliable than individual assessment by surgical experts.

Keywords: computer vision, marker-less video tracking, objective structured assessment of technical skills, objective structured assessment of technical skills, surgical task analysis.

From the ‘Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI; Wisconsin Surgical Outcomes Research (WiSOR) Program, Department of Surgery, University of Wisconsin-Madison, Madison, WI; and ‘Department of Biomedical Engineering, University of Wisconsin-Madison, Madison, WI.

The project described was supported by the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427. Lane Frasier is currently supported by AHRQ F32 HS022403 and also received support from NIH T32 CA90217 and the AAS Research Fellowship Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

The authors report no conflicts of interest.

Reprints: Robert G. Radwin, PhD, University of Wisconsin-Madison 1550 Engineering Drive Madison, WI 53706. E-mail: radwin@wisc.edu.

Copyright © 2017 Wolters Kluwer Health, Inc. All rights reserved.

ISSN: 0003-4932/17/6903-0374 DOI: 10.1097/SLA.0000000000002478

ORIGINAL ARTICLE

Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating

David P. Azari, MS,* Lane L. Frasier, MD,† Sidhra R. Pavuluri Quamme, MD, MS,‡ Caprice C. Greenberg, MD, MPH,*,† Carla M. Pugh, MD, PhD,*,† Jacob A. Greenberg, MD, EdM,† and Robert G. Radwin, PhD*†

Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating

Lane L. Frasier, MD,* ¼ 0.81). Predicted versus panel ratings for tying had slopes from 0.73 to 1, and intercepts from 0.36 to 1.54 (Average $R^2 = 0.81$). Predicted versus panel ratings for tying had slopes from 0.39 to 0.88, and intercepts from 0.79 to 4.36 (Average $R^2 = 0.57$). The mean square error among predicted and expert ratings was consistently less than the mean squared difference among individual expert ratings and the eventual consensus ratings.

Conclusions: The computer algorithm consistently predicted the panel ratings of individual tasks, and were more objective and reliable than individual assessment by surgical experts.

Keywords: computer vision, marker-less video tracking, objective structured assessment of technical skills, objective structured assessment of technical skills, surgical task analysis.

C

Copyright © 2019 Wolters Kluwer Health, Inc. All rights reserved.
extracting kinematic features of tracked hand movements. We created task-specific rating scales and tracked hand motion records to predict subject matter expert ratings of surgical performance for a series of suturing and tying tasks. We hypothesize that the kinematic features of a surgeon’s hand motion measured using marker-less tracking can be used to accurately model subjective performance ratings made by a panel of experts.

METHODS

Participants

This study utilized in-light OR video cameras (Fig. 1) to capture hand motions of 9 surgeons (6 attendings and 3 residents) during 16 surgical cases. Cases included colorectal, complex upper gastrointestinal, hepatobiliary, surgical oncology, transplant, vascular, thoracic, and cardiac operations. Our institutional review board granted ethical approval for recording and analyzing video data for these operations. Written informed consent was obtained from participant surgeons in advance. Cases were recorded during 9 months. Owing to the positioning and limited field of view of the in-light OR camera, no patient details or protected health information were captured. Audio was not recorded.

Video Selection

Recording of approved cases was initiated remotely after the operation began and stored on a secure hospital computer enabled with a video encoder (AXIS video encoder, Axis Communications, Lund, Sweden). We reviewed the videos using Multimedia Video Task Analysis (MVTA) software (Wisconsin Alumni Research Foundation, Madison, WI) developed by Yen and Radwin. MVTA allows real-time review, labeling, and exporting of video data for these operations. Written informed consent was obtained from participant surgeons in advance. Cases were recorded during 9 months. Owing to the positioning and limited field of view of the in-light OR camera, no patient details or protected health information were captured. Audio was not recorded.

Rating Scales

Subjective visual-analog rating scales were created for evaluating surgical performance during short clips (5–30 seconds) for motion economy, fluidity of motion, and tissue handling (Fig. 2). We developed the scales utilizing the OSATS motion scales (ie, respect for tissue, time and motion, and instrument handling) as assessment blueprints. The scales were framed to evaluate performance during a short clip rather than for an individual during the whole procedure (thereby omitting procedure-specific checklists), and defined to comprise the entire range (0–10) of possible behaviors observed during that segment.

Fluidity of motion is a measure of hesitancy, pauses, or changes in direction and “resets,” which may be a component of Moulton’s “slowing down,” or contribute to time spent idle.

Tissue handling quantifies the appropriateness of the surgeon’s force and tension when manipulating the tissue, and varies based on the tissue’s friability and fragility. Motion economy is defined as efficiency of movement, or conservation of energy in any trajectory. Such behavior is consistently documented as a mark of expert psychomotor behavior, and increasingly studied as a measure of surgical skill.

A consensus panel of 3 expert surgeons (CCG, JAG, CMP) viewed each clip in random order and independently rated hand motion across the scales. Each surgeon announced their ratings to the consensus panel of 3 expert surgeons (CCG, JAG, CMP) viewed each clip in random order and independently rated hand motion across the scales. Each surgeon announced their ratings to the
group. Discrepancies were discussed until consensus was achieved for each rating. If absolute consensus was unable to be achieved, the clip was scored according to the majority evaluation.

**Motion Tracking**

Custom video tracking software, developed in one of the authors’ (RGR) lab, was used to trace a region of interest (ROI) on a visible portion of the surgeon’s hands (generally the index finger or thumb) across successive video frames. We previously used this marker-less tracking approach to examine differences in hand motion between attending and resident surgeons, to discriminate between dominant and nondominant hand motion during reduction mammoplasty and to evaluate hand-motion patterns during simulated clinical breast exams. The x-y pixel location of the ROI for each frame (every 1/30 of a second) was recorded. An analyst selected the size and position of the ROI to track the surgeon’s hand for each video clip, and because of occasional occlusions of the hands and changes in lighting, supervised tracking of the ROI, making manual corrections as necessary.

**Calibration**

In-frame visible measurements of the hands were used to calibrate each video clip from pixels to millimeters. We previously used hand dimensions to provide acceptable estimates of hand speed. Observed proximal interphalangeal joint breadth was scaled to the population means of males (23.0 mm) or females (19.9 mm) depending on the sex of the surgeon. The proximal interphalangeal joint breadth was selected because of its small coefficient of variation for males (0.071) and females (0.064) as determined by the US Army. The video was calibrated based on average hand measurements across 3 different frames, and recalibrated for any change in the position of the in-light camera or the surgical field.

**Variable Selection**

The tracked record of the ROI location allowed quantification of instantaneous displacement, speed and acceleration of the surgeon’s hand, and several additional measurements including jerk, spatiotemporal curvature. Jerk is the third derivative of position with respect to time and generally represents how smooth a motion is, whereas the spatiotemporal curvature function is a measure of direction change based on multiple derivatives of the position signal and is used to indicate the number of discrete movements.

A second-order Butterworth filter within empirically observed upper limits of hand frequency for mono-hand tasks (pass band = 0.005–1.000 Hz) was applied to smooth the acceleration and curvature signals and a Fast Fourier Transform (FFT) function was applied to each signal set to isolate any consistently cyclic and repeated motion patterns, a growing avenue of research in surgical dexterity. Moving averages were calculated for the original and smoothed signal types (speed, acceleration, curvature), as well as relative densities to see how much area is consistently and repeatedly traversed by the hand (this can essentially be interpreted as a proxy for path length, a common variable output by platforms such as the ICSAD). These predictor variables are distributed across the variable families shown in Table 1.

**Modeling Process**

We developed a set of linear regression models to test whether the kinematic features could predict the expert ratings across each of the motion scales. The predictor variables were examined for collinear relationships, and subsets of these selected for regression analysis. In selecting variables to initiate the modeling process, we hypothesized that the average speed and consistent locations in the speed signal would give strong predictions of motion economy...
ratings, whereas the peak arrival rate in the acceleration signal would predict fluidity ratings and jerk or pauses would predict tissue handling ratings. In other words, motion economy might correspond to how fast the surgeons are moving in a consistent area, whereas fluidity might relate to how many changes in speed there are, and tissue handling would be sensitive to any sudden, repeated changes in direction. The models were optimized using the Akaike information criterion to balance the prediction against the number of variables. Every consensus rating scale (fluidity of motion, motion economy, tissue handling) was first modeled for each distinct category of surgery (ie, SBW, SBA, SCA, TBW, TSF, TDP). We then examined the combined predictions from suturing tasks and tying tasks for each rating scale (Fig. 3).

**Table 1.** Features Used for Modeling the Consensus Rating Scales

<table>
<thead>
<tr>
<th>Kinematic variable family names (left) and associated descriptions (right). RMS indicates root mean square; SD, standard deviation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary kinematics</td>
</tr>
<tr>
<td>RMS speed and acceleration</td>
</tr>
<tr>
<td>Moving averages</td>
</tr>
<tr>
<td>Peak counts</td>
</tr>
<tr>
<td>Peak variation</td>
</tr>
<tr>
<td>Idle time (%)</td>
</tr>
<tr>
<td>Working area</td>
</tr>
<tr>
<td>Path density (avg, med, SD, max)</td>
</tr>
<tr>
<td>Speed density (avg, med, SD, max)</td>
</tr>
<tr>
<td>Curvature density (avg, med, SD, max)</td>
</tr>
</tbody>
</table>

**Validation**

To assess the internal validity of the prediction models, we compared the sum of squared errors (SSE), a statistical measure of variance, to the leave-one-out predicted residual sum of squares (PRESS) statistic—a common approach in validating models where test data is not available. The SSE measure assessed the overall fit, whereas the PRESS statistic penalizes the model if it depends on outliers. The closer these values are, the better the model matches the current data without relying on outliers, and the better prediction it provides (Table 2). We calculated the root-mean SSE and root-mean PRESS statistics to normalize the comparison between models, and refer to this value as the model’s “error” for ease of reference. Ideal

![Figure 3](https://example.com/figure3.png)
### TABLE 2. Regression Model Summary Statistics and Predictor Variables

<table>
<thead>
<tr>
<th>Task</th>
<th>Fluidity of Motion</th>
<th>Motion Economy</th>
<th>Tissue Handling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pred vs Obs Variables</td>
<td>Pred vs Obs Variables</td>
<td>Pred vs Obs Variables</td>
</tr>
<tr>
<td>Suturing SBW (n = 19)</td>
<td>(m = 0.85) Peak acceleration ((T = 5000)) (R^2 = 0.82)</td>
<td>(m = 0.87) Peak curvature ((T = 0)) (R^2 = 0.85)</td>
<td>(m = 0.82) Peak variance of acceleration ((T = 3000)) (R^2 = 0.84)</td>
</tr>
<tr>
<td>Test</td>
<td>(b = 0.99)</td>
<td>(b = 0.78) ((T = 100))</td>
<td>(b = 0.30) ((T = 0))</td>
</tr>
<tr>
<td></td>
<td>(R^2 = 0.82) Peak smoothed speed ((T = 100)) Peak curvature ((T = 0))</td>
<td>(R^2 = 0.85) Peak acceleration ((T = 6000))</td>
<td>(R^2 = 0.84) Peak speed ((T = 100))</td>
</tr>
<tr>
<td>Suturing SBA (n = 15)</td>
<td>(m = 0.73) Peak acceleration ((T = 1000)) (R^2 = 0.62)</td>
<td>(m = 0.89) Path density of median speed ((T = 1000)) (R^2 = 0.85) Path acc. ((T = 9000)) (R^2 = 0.005)</td>
<td>(m = 0.75) Path density of mean speed ((V = 7)) (R^2 = 0.013) Peak acceleration ((T = 9000)) (R^2 = 0.118)</td>
</tr>
<tr>
<td></td>
<td>(b = 1.43) RMS speed</td>
<td>(b = 1.59) ((R = 7))</td>
<td>(b = 1.54)</td>
</tr>
<tr>
<td></td>
<td>(R^2 = 0.62) Peak rate speed ((T = 1000)) Median speed</td>
<td>(R^2 = 0.85) Peak acc. ((T = 1000))</td>
<td>(R^2 = 0.106) Path density of speed ((T = 100))</td>
</tr>
<tr>
<td>Suturing SCA (n = 10)</td>
<td>(m = 0.75) Weighted avg. of smooth speed ((W = 18s)) (R^2 = 0.63)</td>
<td>(m = 0.85) Path density of max. curvature ((T = 100)) (R^2 = 0.77) Peak variance of speed ((T = 300)) (R^2 = 0.90)</td>
<td>(m = 0.93) Weighted average of smoothed curvature ([0:1]) ((W = 5s)) (R^2 = 0.007) Peak rate of speed ((T = 100)) (R^2 = 0.050)</td>
</tr>
<tr>
<td></td>
<td>(b = 1.13) ((T = 0)) Path density of average curvature</td>
<td>(b = 0.75) ((T = 0))</td>
<td>(b = 0.36)</td>
</tr>
<tr>
<td></td>
<td>(R^2 = 0.63) Path variance of speed ((T = 0))</td>
<td>(R^2 = 0.77)</td>
<td>(R^2 = 0.90)</td>
</tr>
<tr>
<td>Tying TBW (n = 15)</td>
<td>(m = 0.73) Weighted average of acceleration ((W = 70)) (R^2 = 0.65) Fast Fourier transform (FFT) of speed</td>
<td>(m = 0.76) Weighted average of curvature ((T = 80% \text{ of Maximum})) (R^2 = 0.70) Idletime ((T = 50%))</td>
<td>(m = 0.57) Peak variance of speed ((T = 700)) (R^2 = 0.046) Maximum speed (R^2 = 0.092) Peak acceleration ((T = 1000))</td>
</tr>
<tr>
<td></td>
<td>(b = 2.68) RMS speed</td>
<td>(b = 1.65) ((T = 80% \text{ of Maximum}))</td>
<td>(b = 2.88)</td>
</tr>
<tr>
<td>Tying TSF (n = 35)</td>
<td>(m = 0.39) Peak acceleration ((T = 100)) (R^2 = 0.33) Peak speed ((T = 500))</td>
<td>(m = 0.46) Peak acceleration ((T = 9000))</td>
<td>(m = 0.45) Weighted average of smoothed curvature ((W = 5s)) (R^2 = 0.0005)</td>
</tr>
<tr>
<td></td>
<td>(b = 4.36)</td>
<td>(b = 3.96)</td>
<td>(b = 3.83)</td>
</tr>
<tr>
<td>Tying TDP (n = 9)</td>
<td>(m = 0.88) Peak speed ((T = 600)) (R^2 = 0.85)</td>
<td>(m = 0.74) FFT of speed</td>
<td>(m = 0.56) FFT of acceleration</td>
</tr>
<tr>
<td></td>
<td>(b = 0.79) Peak speed ((T = 300)) (R^2 = 0.85)</td>
<td>(b = 1.81) Working area</td>
<td>(b = 3.37)</td>
</tr>
<tr>
<td></td>
<td>(R^2 = 0.85)</td>
<td>(R^2 = 0.65)</td>
<td></td>
</tr>
</tbody>
</table>

Acceptable models (slope between 0.5 and 1.5, with an intercept of 2 or less and \(R^2 > 0.75\)) are indicated in bold text.

Pred = predicted, Obs = observed, \(m = \) slope, with \(\text{predicted} = m \times (\text{observed}) + b\). \(R^2 = \) adjusted coefficient of determination.

Avg. indicates average; Max., maximum; Med., median; R, number of recurrent position traversal as threshold for density calculation; RMS, root mean square; T, threshold in mm/s for speed or mm/s² for acceleration; W, window.
prediction models would have several properties. The predicted and 
ground truth ratings should form a straight line from (0,0) to (10,10), 
such that any unitary increase in expert appraisal would be met with a 
similar predicted increase. We arbitrarily defined acceptable models 
to have a slope between 0.5 and 1.5, an intercept within ±2 of zero, 
and an $R^2 > 0.75$. Although other thresholds could be defined, these 
bounds were chosen to ensure a limited difference between the 
ground truth and the prediction, and a prediction consistent through- 
out the domain of each scale.

RESULTS

Video Data

In total, 103 video clips (mean time = 11.72 seconds) were 
recorded of 6 attending and 3 resident surgeons performing suturing 
and tying tasks throughout 16 varied operations. These included 
SBA, SCA, and SBW suturing clips (n = 44), and TBW, TSF, and 
TDP tying clips (n = 59). The marker-less hand tracking frame-by-
frame position data provided >1500 kinematic variables, spread 
across the variable types in Table 1.

Task Expert Rating Scales

Each clip was observed by 3 expert surgeons and rated 
along 0 to 10 analog scales for motion economy, fluidity of 
motion, and tissue handling. The expert raters achieved consensus 
after a maximum of 3 iterations for all the video clips. Observed 
ratings for suturing tasks ranged from 2 to 9 (mean $= 5.91$, SD $= 1.62$). Tying task ratings ranged from 3 to 9 (mean $= 7.12$, SD $= 1.10$) and were particularly skewed toward the upper end of the scale.

Prediction Models of Expert Ratings

The linear regression slopes and intercepts for each of the 
rating scales and task types are seen in Table 2. The models for 
fluidity of motion for suturing (Fig. 3A; slope $= 0.86$, intercept $= 0.80$, $R^2 = 0.86$) and motion economy for suturing (Fig. 3B; slope $= 0.89$, intercept $= 0.64$, $R^2 = 0.88$) had the best predictions, and moderately better than tissue handling of suturing (Fig. 3C; slope $= 0.76$, intercept $= 1.1$, $R^2 = 0.69$).

The predicted versus consensus ratings of motion economy for 
tying tasks (Fig. 3E; slope $= 0.65$, intercept $= 2.51$, $R^2 = 0.64$) had better predictions than tissue handling (Fig. 3F; slope $= 0.53$, intercept $= 3.33$, $R^2 = 0.52$) and fluidity of motion (Fig. 3D; slope $= 0.54$, intercept $= 3.28$, $R^2 = 0.54$).

Prediction Model Validity

Linear prediction models were validated by comparing their 
difference, or “error,” between root-mean predicted residual (PRESS) and root-mean sum of square errors (SSE). Motion econo-
my had the least error (0.17, 0.07) during bowel anastomosis and 
superficial tying, respectively, as well as the least average error (0.27) 
for suturing ratings overall. Fluidity of motion models similarly 
performed well for suturing and tying along the body wall with errors 
of 0.21 and 0.11. Tissue handling had the highest error for both 
suturing and tying tasks (0.52, 0.26). Low errors suggest strong 
positive relationships between hand kinematics and expert ratings—a 
necessary component of Kane and Messick’s modern approach to 
validity.

DISCUSSION

This study demonstrates that computer vision capturing kinematic 
data from marker-less motion tracking of video records offers 
an objective and scalable approach for measuring surgical skill, and 
establishes evidence consistent with Kane’s validity framework. 
We created subjective rating scales for fluidity of motion, motion 
economy, and tissue handling using existing OSATS measures as an 
assessment blueprint. We used kinematic features of the hands to 
develop prediction models of the expert ratings and compared the 
model performance against the variance among experts. Models 
consistently had less variance than the individual experts exhibited 
before consensus (Fig. 4). In this process, we identified the kinematic 
measures most closely linked with expert surgeons’ ratings of 
common surgical tasks. Because this approach relies only on noninvasive video tracking and not invasive markers for motion capture,
represents a critical step forward in a scalable and reproducible method for objective assessment of surgical technical skill during actual, open operations.

Numerous features were extracted from the video. Hand position, speed, acceleration, and curvature of motion were measured and recorded. Based on previous expectations of psychomotor performance, a series of additional variables were calculated, including peak frequency and variance, path density, working area, and moving average. Peak arrival rates in both the unfiltered and smoothed speed and acceleration signals were consistently correlated and significant in the prediction models.

The fluidity of motion and motion economy models for all suturing tasks had slopes between 0.73 and 1, and intercepts between 0.30 and 1.54. Although models of tissue handling underperformed the former, the tissue handling prediction for complex or bowel anastomosis suturing tasks had a slope of 0.93 and intercept of 0.36. Many models utilized the peak rate arrivals in the acceleration signal, and achieved slopes close to 1, with intercepts between 0.5 and 1.5, which indicates excellent fit. These results support the conclusion that kinematic features of a surgeon’s hand motion, measured using marker-less tracking, can accurately model subjective performance ratings made by a panel of experts. Such measures are useful in providing feedback and automatic feedback, and necessary for developing evidence of validity for future competency-based assessments.

There are several limitations to consider. First, this study was contingent on access to video of live operations. Thus, not all scores were observed across all task subtypes. Tying tasks demonstrated skewed expert ratings or had a limited range of scores, reducing the variance and making these tasks more difficult to predict, despite low error estimates. Additionally, the task-specific rating scales themselves did not always address the surgical context. For example, a surgeon with excellent fluidity may exhibit periods of frequent pauses to gather information, avoid distraction, compensate for a hand tremor, or simply resolve confusion or nervousness. Although the detection itself does not suggest a particular decision or course of action to improve, automatically identifying events such as “slowing down,” as Moulton describes, could streamline video analysis to target critical points in an operation and provide insight into surgical decision making. In other words, computer vision technology may help automatically detect difficulties in an operation, demonstrated by acute changes in hand motion.

This technology also ignores the consequences to a procedure or a patient, and would need to meet additional evidence requirements outlined by Kane and Messick before any deployable assessment routine takes shape. Such assessment would also need to incorporate whether the surgeon identifies the anatomy and adjusts their technique appropriately, commits errors, or generally performs the operation correctly. A suturing task with low fluidity of motion (high hesitancy, frequent pauses) may exhibit significant motion while repositioning the driver multiple times, without contacting tissue, despite a continuous motion tracking record. It is possible that such events may be isolated by combining the different scale predictions (ie, sudden poor tissue handling during a period of low fluidity) and high motion economy may hold valuable information about the current state of the surgery.) Future work may address these challenges by considering the implications of unfiltered and moving average of assessment scores for a whole surgical case. These investigations may consider the quality of observed ties or stitches, and focus on more advanced relationships between kinematic features and the overall surgical state, utilizing patterns of motion across repeated cycles and language models to automatically classify surgeons and procedures.

We anticipated that fluidity would produce the best prediction outcomes based on recorded speed, and that tissue handling would be most difficult to model, given dependence on tissue type. Similarly, the more complex anastomoses and tying tasks were expected to have greater variability in hand movements—potentially making them more difficult to predict. The tissue handling model for anastomoses (bowel and complex) performed better than tissue handling along the body wall, supporting the hypothesis that tissue handling ratings may be more sensitive to the categorization process and exhibit higher variation. It is also possible that the cues the raters used to gauge tissue handling were not as readily identifiable in the motion tracking record for suturing along the body wall (ie, hand shape, individual finger dexterity), and therefore currently unavailable to the automatic routine.

Arranging time for the expert panel to view and rate the videos was the most difficult aspect of the study. Surgeons reported achieving consensus relatively easily after discussion, but had general difficulty when the field of view was limited. How surgeons resolved disagreements for these clips may suggest additional contextual features to improve the predictions, and may help to establish evidence of validity in the response process itself. Collecting and applying motion-tracking routines to the videos played in real-time was also resource-intensive and time-consuming. More complex procedures with multiple changes in patient or camera positions, shifting inclusion automatic feedback, and necessary for developing evidence of validity for future competency-based assessments.

Azari et al

In future studies, the overall difficulty with skewed distributions could be mitigated by collecting videos of specific clinically simulated scenarios, wherein variety and experience are intentionally controlled to accommodate the full range of performance ratings. In this setting, the rating scales and kinematics could incorporate different surgical approaches, as some surgeons throw 1-handed ties, whereas others prefer 2-handed ties and the kinematic path for these are quite different, limiting the features available for comparison. An expert viewer may be able to distinguish between a successfully thrown knot and a dropped ligature, but depending on how the surgeon moves after the error, their kinematic path may look identical. Such situations continue to require subjective supervisory intervention and assessment, and future work is needed to automatically identify these contextual factors, and assess measures of reproducibility among different stations and raters in more controlled (ie, benchtop) simulations.

Access to immediate, reproducible kinematic-based feedback based on video review can inform self-assessment, direct practice of specific tasks, and build overall surgical skill. This may help to provide a venue in which skill development is quantifiably traceable to training interventions. If used to track expert surgeons, the capacity to deconstruct surgical skill can provide a deeper understanding of the kinematics of surgical performance and aid in the development of novel approaches to skill acquisition. Understanding the components of performance and providing such feedback can potentially shorten the current learning curve and help detect skill decay using objective measures either during periods of inactivity or toward the end of surgical careers.

The findings in this study represent a measurable step forward in creating more objective, reproducible, and accessible assessments of surgical skill commensurate with direct video observation by panel of expert raters. The prediction models have the potential to be packaged into automatic, on-demand feedback in and out of training.
settings, providing a reliable measure of assessment and consistent feedback to facilitate direct practice of hand motions for specific tasks. Future work is needed to explore the capacity of motion capture alone to make decisions about trainee competence. As video capture is becoming increasingly common in the operating room, computer vision motion tracking allows for a uniquely scalable approach to surgical skill analysis.

REFERENCES

41. Hogan N, Sterndad D. Sensitivity of smoothness measures to movement duration, amplitude, and arrests, J Mot Behav [Internet]. 2009;41:529–534.