Visceral Medicine

Research Article

Visc Med 2020;36:463–470
DOI: 10.1159/000512437

Received: June 9, 2020
Accepted: October 20, 2020
Published online: October 28, 2020

How to Bring Surgery to the Next Level: Interpretable Skills Assessment in Robotic-Assisted Surgery

Kristen C. Brown   Kiran D. Bhattacharyya   Sue Kulason   Aneeq Zia
Anthony Jarc

Advanced Product Development, Intuitive Surgical, Inc., Norcross, GA, USA

Keywords
Surgical skill evaluation · Training · Robotic surgery · Objective performance indicators · Education

Abstract

Introduction: A surgeon's technical skills are an important factor in delivering optimal patient care. Most existing methods to estimate technical skills remain subjective and resource intensive. Robotic-assisted surgery (RAS) provides a unique opportunity to develop objective metrics using key elements of intraoperative surgeon behavior which can be captured unobtrusively, such as instrument positions and button presses. Recent studies have shown that objective metrics based on these data (referred to as objective performance indicators [OPIs]) correlate to select clinical outcomes during robotic-assisted radical prostatectomy. However, the current OPIs remain difficult to interpret directly and, therefore, to use within structured feedback to improve surgical efficiencies. ***Methods:*** We analyzed kinematic and event data from da Vinci surgical systems (Intuitive Surgical, Inc., Sunnyvale, CA, USA) to calculate values that can summarize the use of robotic instruments, referred to as OPIs. These indicators were mapped to broader technical skill categories of established training protocols. A data-driven approach was then applied to further sub-select OPIs that distinguish skill for each technical skill category within each training task. This subset of OPIs was used to build a set of logistic regression classifiers that predict the probability of expertise in that skill to identify targeted improvement and practice. The final, proposed feedback using OPIs was based on the coefficients of the logistic regression model to highlight specific actions that can be taken to improve. ***Results:*** We determine that for the majority of skills, only a small subset of OPIs (2–10) are required to achieve the highest model accuracies (80–95%) for estimating technical skills within clinical-like tasks on a porcine model. The majority of the skill models have similar accuracy as models predicting overall expertise for a task (80–98%). Skill models can divide a prediction into interpretable categories for simpler, targeted feedback. ***Conclusion:*** We define and validate a methodology to create interpretable metrics for key technical skills during clinical-like tasks when performing RAS. Using this framework for evaluating technical skills, we believe that surgical trainees can better understand both what can be improved and how to improve.

Introduction

There is increasing evidence that surgeon technical skills influence postoperative patient outcomes [1–7]. Methods to estimate technical skills and provide feedback throughout surgeons' learning curves can lead to more efficient training [1, 8–10]. In turn, this could lead to improved patient outcomes. However, the challenge lies not only in developing the proper methods to measure technical skills accurately, but also in delivering feedback that is understandable and, thereby, actionable to improve performance.

Expert-evaluated methods can provide great insight into a subset of surgeries but scale poorly due to the time-

karger@karger.com
www.karger.com/vis

Kristen C. Brown
Advanced Product Development
Intuitive Surgical, Inc., 5655 Spalding Drive
Norcross, GA 30092 (USA)
Kristen.Brown2@intusurg.com

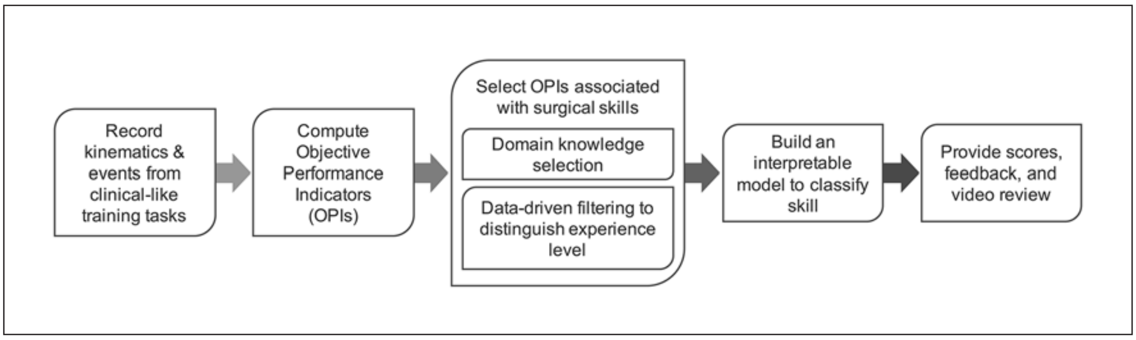


Fig. 1. An overview of the workflow to provide data-driven, structured feedback to surgical trainees.

consuming process of human rating [11]. In-person expert evaluation remains subjective in nature and can result in varying feedback from different experts. Furthermore, evaluation of surgeon performance from only a small subset of cases can be confounded by case-to-case fluctuations of patient disease factors, the composition of the OR team [12], or days off between cases [13] (to name just a few). For these reasons, expert-evaluated methods may have limited insight into which intraoperative factors influence postoperative outcomes. Surrogate measures of surgical skill and performance, such as surgeon experience or level of training, while correlated to patient outcomes [5, 7] and case-by-case evaluations of skill [3], do not provide the necessary information for deliberate practice and improvement.

In comparison, computer-aided methods of analyzing surgical movements and tool use could scale well across many surgeries since these techniques are more amenable to automation. Moreover, computer-aided analysis of kinematic and event data shows promise in providing objective evaluation of surgical skill both during training exercises [14–18] and during surgery [19, 20], even achieving correlation to postoperative patient outcomes [1, 3, 4]. Prior evaluations of surgical skill have focused on producing a categorical label, such as expert or novice, or reproducing a subjective score, like GEARS, over entire procedures or tasks. While computer-aided methods still require manual annotation of procedural steps or tasks, recent advances to recognize clinically relevant surgical activities using machine learning [21–26] suggest that annotation of tasks within surgeries and consequent analyses of kinematics could be entirely computer automated [27]. Even though computer-aided methods have the potential to significantly reduce the time needed to evaluate surgical performance, the metrics produced by this assessment can be difficult to interpret directly unless appropriately designed.

In this study, we propose a new framework (Fig. 1) to evaluate technical skills learned during training activities on a porcine model through interpretable metrics. We set out to explore the feasibility of breaking tasks into technical skills, while still capturing objective information that distinguishes expertise. We hypothesized that models distinguishing expertise built on a subset of objective performance indicators (OPIs; grouped by technical skill) would perform as well as models built on all OPIs and that such models can provide more interpretable feedback to trainees.

Methodology

Participants

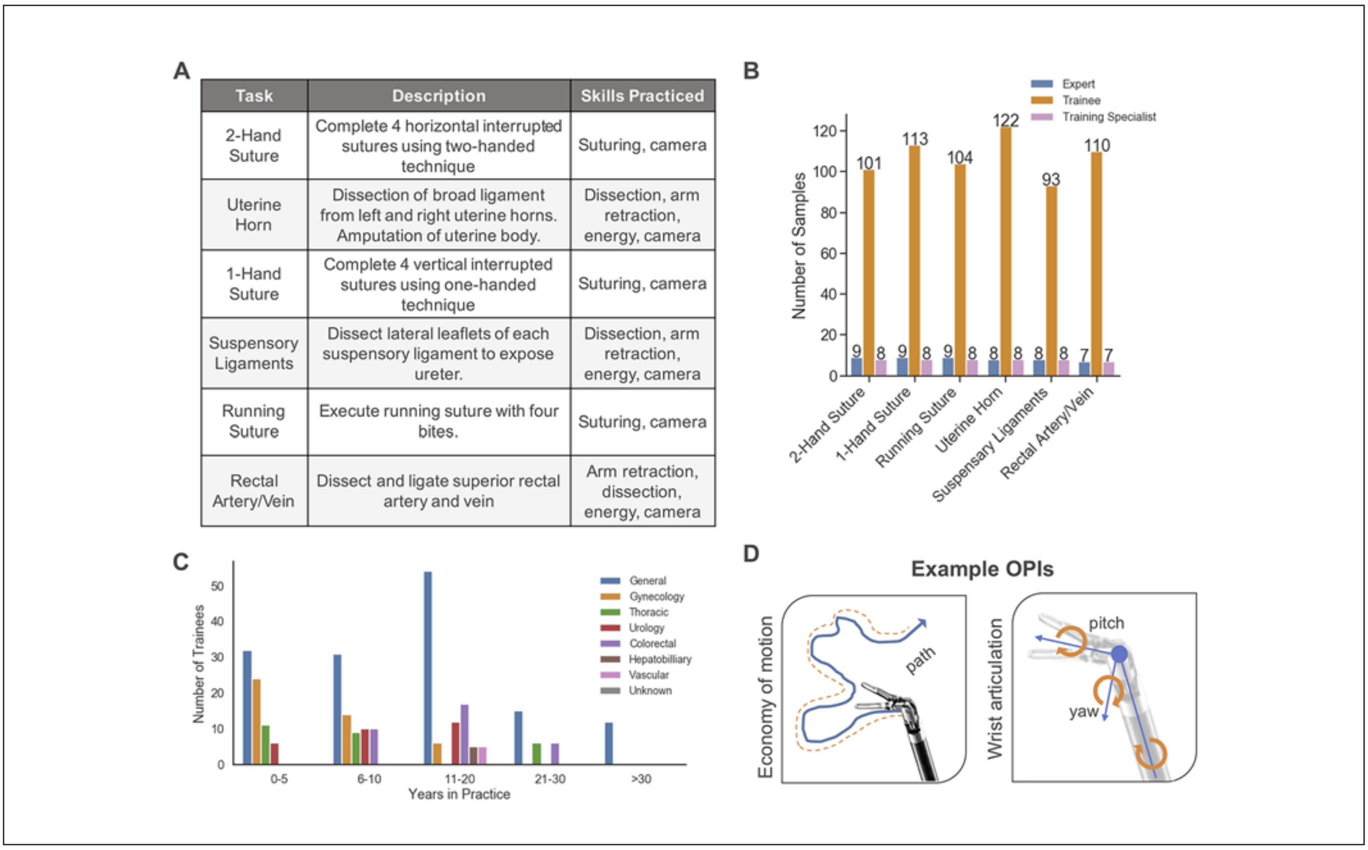
Each participant performed clinical-like tasks that were designed to practice technical skills in using the robotic platform in a porcine model described in Figure 2A. Participants were split into 3 categories of users of the da Vinci surgical system: (1) *trainee*, (2) *expert surgeon*, and (3) *training specialist* (Fig. 2B). Trainees were surgeons that do not have robotic surgery experience but have varying years of non-robotic experience across several specialties (Fig. 2C). Expert surgeons performed >1,000 da Vinci robotic procedures. Training specialists were non-surgeon, expert users that are experienced in the assessed training exercises with ~300–600 h of practice on or use of robotic platforms. Training specialists specialize in basic robotic technical skills targeted in these introductory exercises. They train hundreds of new robotic surgeons in this introductory course.

Data Collection

We recorded synchronized video, kinematic (instrument positions and joint angles), and event (button presses, etc.) data from the da Vinci surgical systems. Each participant contributed a single recording. The start and stop times of each task were labeled to associate the kinematic and event data to the correct tasks. There were 7–9 tasks from expert surgeons and training specialists and 93–122 from the trainee group (Fig. 2B).

Objective Performance Indicators

OPIs were computed to summarize the complex time series data into meaningful information. We calculated 43 OPIs (see online suppl. Table for a complete list; for all online suppl. material, see www.karger.com/doi/10.1159/000512437) that estimate technical efficiencies when using the robotic platform (for examples, see Fig. 2D). We developed OPIs with interpretability in mind, and those too difficult to explain were not included. Similar metrics have been explored to characterize technical skills in inanimate



Task	Description	Skills Practiced
2-Hand Suture	Complete 4 horizontal interrupted sutures using two-handed technique	Suturing, camera
Uterine Horn	Dissection of broad ligament from left and right uterine horns. Amputation of uterine body.	Dissection, arm retraction, energy, camera
1-Hand Suture	Complete 4 vertical interrupted sutures using one-handed technique	Suturing, camera
Suspensory Ligaments	Dissect lateral leaflets of each suspensory ligament to expose ureter.	Dissection, arm retraction, energy, camera
Running Suture	Execute running suture with four bites.	Suturing, camera
Rectal Artery/Vein	Dissect and ligate superior rectal artery and vein	Arm retraction, dissection, energy, camera

Fig. 2. Data, participants, and methodology. **A** Table of each task in the study, along with a description and skills practiced from the training protocol. **B** Number of recordings for each clinical-like training task grouped by experience level. The number of recordings containing each task per group is labeled on each bar. **C** Trainee backgrounds in this study described as years of non-robotic experience in practice and divided by specialty. **D** Two example OPIs calculated for this study. Economy of motion represents total distance traveled along a linear path. Wrist articulation is measured through tracking 3 joint angle movements. A full list and descriptions can be found in the online supplementary material.

models, virtual reality, and surgical settings (please see [28, 29] for reviews). For preliminary analysis of OPIs among participants, we compared training specialists and expert surgeons through cluster analysis, Wilcoxon rank sum test for significance, and visualization of task duration distributions. For cluster analysis, all OPIs with non-zero variation are first converted to a z-score, followed by calculating hierarchical clusters based on Euclidean distances.

Technical Skills Mapping

To provide targeted feedback, we defined a set of broader technical skill categories that can be evaluated separately within the same tasks (Fig. 2A, "Skills Practiced") based on established training protocols. Each OPI is mapped into a technical skill category prior to the data-driven selection based on whether or not the instrument or event is used in each category during the task. For example, while camera control activation may be important in overall expertise, it is not directly related to the trainee's skill in using energy.

Experimental Design

To estimate technical skills, we compared OPIs of trainee surgeons to an experienced group (expert surgeons and training specialists). For each task, we used these groups to classify technical skill based on OPI values. In order to evaluate our hypothesis, we created 2 sets of models and evaluated their performances: (1) task models that include all 43 OPIs and (2) models specific to each skill evaluated in a task using OPIs from the technical skills mapping. Given the large number of trainee participants (Fig. 2B), we held back 5 randomly selected trainees per task from this process to demonstrate the model prediction to feedback process, leaving 88–117 for feature selection. Both model sets were used to predict the probability of expertise for 5 trainee recordings that were held back from model building.

Cross-Validation and Data Resampling

We used a nested 5-fold cross-validation approach described in Figure 3A to obtain a more robust measure of classifier accuracy. First, the data is split pseudo-randomly into training and test sets (80/20%). By using a stratified sampling approach, we maintain the original class distributions. The training data per fold is then used to perform the inner 5-fold cross-validation of recursive feature elimination (RFE), requiring an additional train-test split. This approach is done for 2 sets of models: (1) an overall task model using all 43 OPIs and (2) a skill-specific model using mapped OPIs. We combined upsampling and downsampling approaches to balance sample sizes for each training fold of inner and outer cross-validation loops. We upsampled the *experienced* group by 3-fold using the Synthetic Minority Over-Sampling Technique (SMOTE) [30]. The trainee data was downsampled in 2-steps. First, we used the Neighborhood Cleaning Rule [30, 31] to reduce

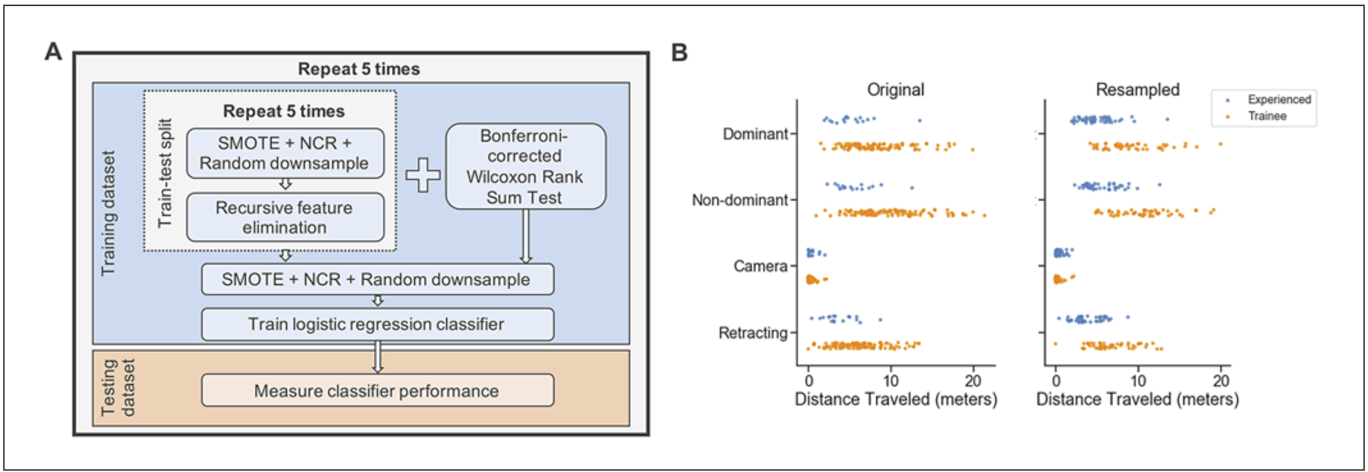


Fig. 3. A The nested cross-validation method for feature selection and model validation described in the Methodology. **B** A dot plot of economy of motion values for all 4 instruments in the Uterine Horn task before (left) and after (right) the resampling approach described in the Methodology. Experienced group contains both expert surgeons and non-surgeon training specialists.

noisy data of the trainee group by removing points far from other trainee samples. We then applied random undersampling of remaining samples to balance the groups to an approximate 50/50 ratio [30]. When the class sizes are small, these combination methods tend to perform better than their random counterparts [32]. An example of OPI data before and after this process is shown in Figure 3B. This is done for both cross-validation loops to avoid performing feature selection only on synthetic samples.

OPI Selection

Within each training task and skill combination, RFE with logistic regression as a base estimator was used to determine the features (OPIs) to build a model. To avoid including OPIs with spurious relationships to expertise, we used 2 methods to rank OPIs: (1) the Wilcoxon rank sum test using a Bonferroni-adjusted p value cutoff at $p < 0.05$ [33] and (2) RFE with logistic regression as a base-estimator [34]. The OPIs selected in each method are combined. Often, these ensemble methods for feature selection are used to provide better feature sets [35].

Wilcoxon rank sum tests were not performed on resampled data in the inner cross-validation loop alongside RFE to avoid biasing p values. RFE selects features by recursively considering smaller and smaller sets of features until a minimum number is reached, in this case 1, ranking each OPI by importance [34, 36]. At each RFE step, the model is evaluated on the test data, which is not resampled, to determine the minimal and optimal number of OPIs. The final feature set is the union of OPIs selected by RFE and Wilcoxon.

Skill Classification Model

After OPI selection, the training data is used to build a logistic regression classifier to predict experience level. The model is then tested on the test data, which was not subject to resampling. This is repeated for the other 4 folds to obtain a cross-validated estimate of each model's performance. We measured performance using 2 metrics: a balanced accuracy score (average of recall between both groups) to account for the large differences in sample sizes [37] and Matthews correlation coefficient (MCC). MCC is a balanced quality measure of classification, ranging from –1 (poor prediction) to 1 (perfect prediction) [38, 39]. The coefficients of each logistic regression model are converted to odds ratios to determine how the model uses OPIs to classify surgeon skill. Based on the directionality and magnitude of odds ratio values, feedback can target improvement among the reported OPIs.

Results

Analysis of Participant Groups

We analyzed OPIs of training specialists and expert surgeons to determine if there were significant differences in these groups of highly skilled users. For each task, we found no significantly different OPIs ($p < 0.05$) after multiple testing corrections. Hierarchical clustering of the training specialists' and experienced surgeons' data shows that the 2 groups do not form distinct groups based on their OPIs (Fig. 4A). A commonly used metric for distinguishing surgical skill is task completion time, which has a similar distribution across tasks for training specialists and experienced surgeons but that does differ from the trainee group (Fig. 4B). Expert surgeon and non-surgeon training specialists were combined into a single, high-skilled grouped deemed as the experienced group to classify technical skills.

Data-Driven OPI Selection

Each skill-task combination of mapped OPIs started with 4–20 OPIs and RFE reduced it to 1–18 OPIs. RFE for task models displayed high balanced accuracies (cross-validated [CV] score; Fig. 4C) that typically plateau early. Wilcoxon testing produced overlapping feature sets and added 0–11 OPIs to the final models.

Classification Using Broad Technical Skill Models

The balanced accuracy and MCC of each task-specific model is reported in Figure 5A, along with the same met-

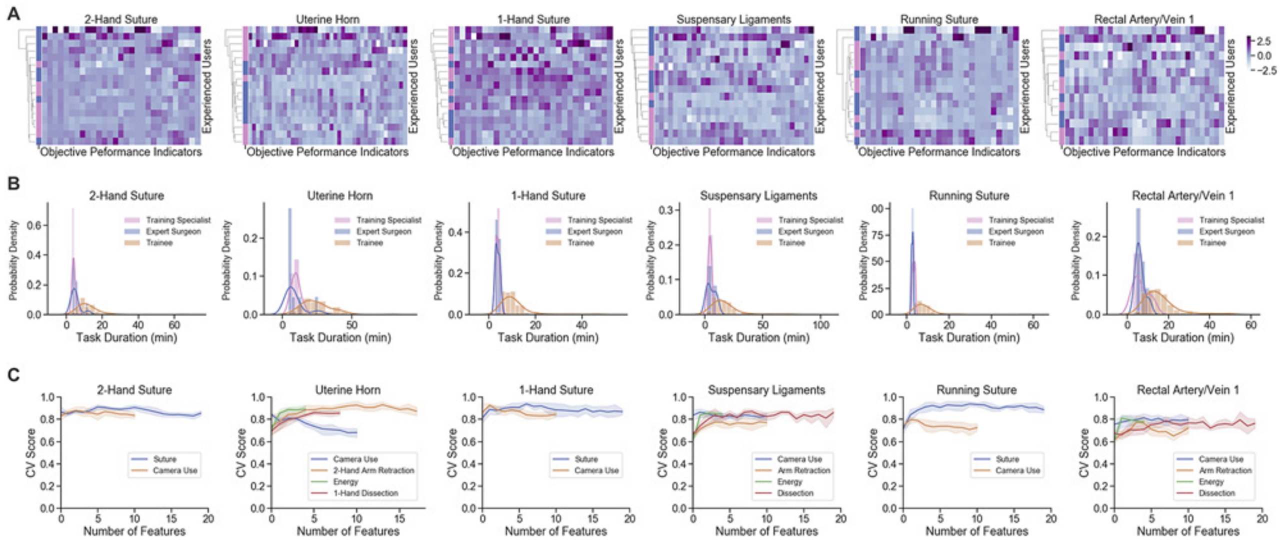


Fig. 4. Analysis of OPIs across experience levels and feature selection. **A** Dendrograms of hierarchical clustering results for expert surgeons (blue) and non-surgeon training specialists (pink) along the *y*-axis. Heatmaps display *z*-score normalized values for OPIs and task durations with non-zero variance along the *x*-axis. **B** Distributions of task durations split by experience level. **C** Cross-validated (CV) scores (balanced accuracy, *y*-axis) of varying numbers of OPIs per skill (*x*-axis) using RFE.

A

Task	Accuracy[1]	MCC[2]	Skill	Accuracy[1]	MCC[2]
2-Hand Suture	89.28	0.78	Suture	90.54	0.74
			Camera Use	86.34	0.69
Uterine Horn	93.14	0.83	1-Hand Dissection	84.58	0.63
			2-Hand Arm Retraction	87.25	0.70
			Energy	84.91	0.59
			Camera Use	76.32	0.49
1-Hand Suture	95.67	0.90	Suture	83.83	0.63
			Camera Use	95.44	0.87
Suspensory Ligaments	91.21	0.75	Dissection	77.43	0.51
			Arm Retraction	78.62	0.45
			Energy	85.72	0.64
			Camera Use	89.57	0.79
Running Suture	98.27	0.91	Suture	94.44	0.86
			Camera Use	77.40	0.53
Rectal Artery/Vein	80.24	0.52	Dissection	78.33	0.50
			Arm Retraction	80.48	0.46
			Energy	83.33	0.65
			Camera Use	79.29	0.52

[1]Balanced accuracy, cross-validation average
[2]Matthews Correlation Coefficient, cross-validation average

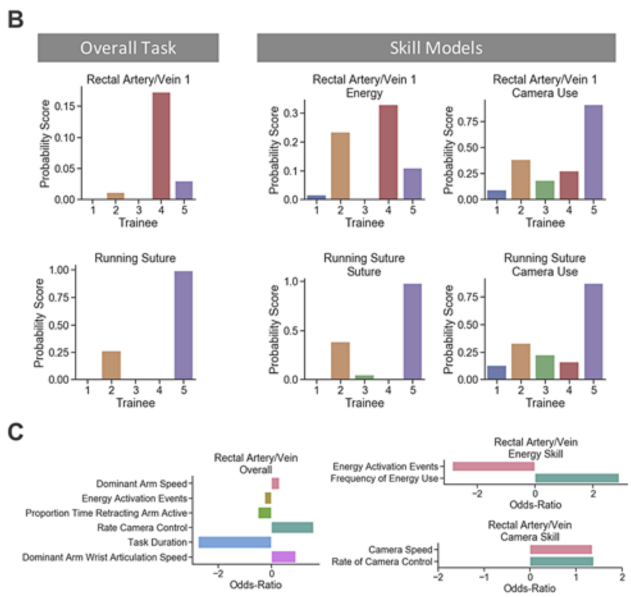


Fig. 5. Task- and skill-based model performance. **A** Table of average cross-validation performance metrics, balanced accuracy, and Matthews correlation coefficient (MCC) for each of the skill-task and overall task logistic regression models. **B** Example of predicted probabilities for 5 trainees by overall task models and 2 skill-mapped models for Rectal Artery/Vein and Running Suture. **C** Odds ratio for Rectal Artery/Vein task model and skill-mapped models for energy and camera use.

rics for each skill-based model per task. Each task model accuracy ranged from 80.24 to 98.27% and MCCs from 0.52 to 0.78. Skill model accuracies ranged from 76.32 to 95.44% and MCCs from 0.45 to 0.87. The skill models produce performance metrics mostly no lower than 10% of each task model's balanced accuracy (14/18), with no one task standing out. Thus, splitting each scoring model into skills maintains similar accuracy to more commonly used scoring models that would consider all OPIs together. While the MCC was well above 0 across models, they were less consistent by the same criteria (8/18).

Interpreting Predictions and Providing Feedback

For 5 trainee predictions (Fig. 5B, C), the Rectal Artery/Vein task shows low probabilities, scoring correctly as beginners, with the highest for Trainee 4. For running suture, there is 1 trainee with a high probability of expertise. Two corresponding skill models for each task present a wider range of probabilities that are more consistent with task scores, like Trainee 2 or 5 for running suture, or differ, like Trainee 3 or 5 for Rectal Artery/Vein (Fig. 5B). The overall Rectal Artery/Vein task model includes only 1 OPI for energy and camera use (Fig. 5C, Rectal Artery/Vein). The inclusion of task duration may have improved skill models overall, as this is a commonly used metric for classifying surgical skill, but not specific to a single technical skill. Skill models included frequency of energy use and median camera speeds, while task models did not. Overall task feedback could include primarily improving the task completion time, increasing the frequency of camera control, and practicing increasing the speed of the dominant arm. Energy skill feedback would include practicing reducing unnecessary energy activation (reduce total events), while applying energy more frequently in shorter time periods (increase frequency). Camera skill feedback would include not only increasing the frequency of adjusting the camera to improve the field of view, but also doing so at faster speeds, which may relate to inefficiencies in camera adjustments.

Discussion

Early evaluation of robotic skill can be broken down into specialty-agnostic skill categories, with experienced non-surgeon experts displaying similar technical proficiencies as expert surgeons in a structured training setting. This evaluation is built on the assumption that experienced surgeons and non-surgeon, expert users are highly skilled in these activities, while new robotic-assisted surgeons are not. While this may often be true, it can lead to noisy data, as some new surgeons may become adept at certain technical skills quickly. Perhaps, a better model would be based on evaluation of performance by expert surgeons, but even that has its caveats, such as challenges with inter-rater reliability [40–42].

The probabilities computed from the proposed models can be presented as a score either raw, ranked, or reweighted to give the trainee a sense for where they can improve. Skill models may give more insight into where each trainee needs to focus, rather than a single score that may be driven by different OPIs. Feedback from an overall task model would generally indicate task-specific expertise, recommending practice in tasks like Rectal Artery/Vein, which may not be useful to all specialties. From a task model alone, we might suggest a trainee increase the rate of camera control activation but, without the other camera OPIs, we miss out on recommendations on how to control the camera to improve. We propose that broad skill categories give interpretable guidance by first breaking each task into technical skills that are common across specialties, but this has yet to be validated. Instead of informing the surgeon of their poor performance of Rectal Artery/Vein, they are told where their inefficiencies lie in camera, energy, dissection, and arm retraction skills. Examining further the feedback for each skill, the model can be used to suggest improvement with more specific context, such as inanimate training or virtual reality simulation that specifically target these smaller, technical proficiencies.

Additionally, the scores and feedback can be paired with videos of a trainee's task performance alongside a video of a peer or experienced surgeon for comparison(s). This allows a trainee to review their own performance, taking care to reflect on specific skills within the video and how the other surgeons perform the same task. Our framework does not encompass all technical skills, such as force sensitivity [43], or aspects of clinical judgment [44], which surgeons seem to be able to derive from video review.

The objective evaluation of surgical technical skill using computer-aided methods is an area of growing interest in the broader surgical education community to address challenges with subjective rating scales [43, 45]. Our work explores a quantitative method for providing more interpretable guidance for surgical trainees early in their learning curves by utilizing a logistic regression classifier built on skill-mapped OPIs. This presents a potential advantage over earlier work because we may be able to integrate better into the surgeon training pathways by targeting broader technical skills within clinical-like tasks (as opposed to dry-lab activities [45]) and specific deliberate practice recommendations. The utility and advantage of the work in surgeon training remains to be evaluated in future studies; however, we believe our framework can be applied to clinical settings to enable continued feedback throughout a surgeon's learning curve.

Acknowledgements

We would like to thank Jay Rhode for providing guidance in developing skill definitions from a clinical perspective and other helpful discussions. We would also like to thank Linlin Zhou for technical assistance in processing the data recordings and helpful discussions.

Statement of Ethics

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. All applicable international, national, and/or institutional guidelines for the care and use of animals were followed. Informed consent was obtained from all individual participants included in the study.

Conflict of Interest Statement

Kristen C. Brown, Kiran D. Bhattacharyya, Aneeq Zia, Sue Kulason, and Anthony Jarc are employees of Intuitive Surgical, Inc.

Funding Sources

There were no funding sources.

Author Contributions

K.C.B. contributed to the design, analysis, and interpretation of data; drafted and revised the work; approved the final version; and agrees to be accountable for all aspects of the work. K.D.B. contributed to the design, analysis, and interpretation of data; drafted and revised the work; approved the final version; and agrees to be accountable for all aspects of the work. A.Z. contributed to the design and interpretation of data; drafted and revised the work; approved the final version; and agrees to be accountable for all aspects of the work. S.K. contributed to the design and analysis of data; drafted the work; approved the final version; and agrees to be accountable for all aspects of the work. A.J. contributed to the conception, design, acquisition, and interpretation of data; drafted and revised the work; approved the final version; and agrees to be accountable for all aspects of the work.

References

1 Chen A, Ghodoussipour S, Titus MB, Nguyen JH, Chen J, Ma R, et al. Comparison of clinical outcomes and automated performance metrics in robot- assisted radical prostatectomy with and without trainee involvement. World J Urol. 2020 Jul;38(7):1615–21.
2 Fecso AB, Kuzulugil SS, Babaoglu C, Bener AB, Grantcharov TP. Relationship between intraoperative non-technical performance and technical events in bariatric surgery. Br J Surg. 2018 Jul;105(8):1044–50.
3 Hung AJ, Oh PJ, Chen J, Ghodoussipour S, Lane C, Jarc A, et al. Experts vs super-experts: differences in automated performance metrics and clinical outcomes for robot-assisted radical prostatectomy. BJU Int. 2019 May; 123(5):861–8.
4 Hung AJ, Chen J, Ghodoussipour S, Oh PJ, Liu Z, Nguyen J, et al. A deep-learning model using automated performance metrics and clinical features to predict urinary continence recovery after robot-assisted radical prostatectomy. BJU Int. 2019 Sep;124(3):487–95.
5 Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, et al.; Michigan Bariatric Surgery Collaborative. Surgical skill and complication rates after bariatric surgery. N Engl J Med. 2013 Oct;369(15):1434–42.
6 Goldenberg MG, Lee JY, Kwong JC, Grantcharov TP, Costello A. Implementing assessments of robot-assisted technical skill in urological education: a systematic review and synthesis of the validity evidence. BJU Int. 2018 Sep;122(3):501–19.
7 Chen J, Chu T, Ghodoussipour S, Bowman S, Patel H, King K, et al. Effect of surgeon experience and bony pelvic dimensions on surgical performance and patient outcomes in robot-assisted radical prostatectomy. BJU Int. 2019 Nov;124(5):828–35.
8 Vedula SS, Ishii M, Hager GD. Objective assessment of surgical technical skill and competency in the operating room. Annu Rev Biomed Eng. 2017 Jun;19(1):301–25.
9 Azari D, Greenberg C, Pugh C, Wiegmann D, Radwin R. In search of characterizing surgical skill. J Surg Educ. 2019 Sep-Oct;76(5):1348–63.
10 Reiley CE, Lin HC, Yuh DD, Hager GD. Review of methods for objective surgical skill evaluation. Surg Endosc. 2011 Feb;25(2):356–66.
11 Özdemir-van Brunschot DMD, Warlé MC, van der Jagt MF, Grutters JPC, van Horne SBCE, Kloke HJ, et al. Surgical team composition has a major impact on effectiveness and costs in laparoscopic donor nephrectomy. World J Urol. 2015 May;33(5):733–41.
12 Pearce SM, Pariser JJ, Patel SG, Anderson BB, Eggener SE, Zagaja GP. The impact of days off between cases on perioperative outcomes for robotic-assisted laparoscopic prostatectomy. World J Urol. 2016 Feb;34(2):269–74.
13 Curry M, Malpani A, Li R, Tantillo T, Jog A, Blanco R, et al. Objective assessment in residency-based training for transoral robotic surgery. Laryngoscope. 2012 Oct;122(10): 2184–92.
14 Estrada S, Duran C, Schulz D, Bismuth J, Byrne MD, O'Malley MK. Smoothness of surgical tool tip motion correlates to skill in endovascular tasks. IEEE Trans Hum Mach Syst. 2016;46(5):647–59.
15 Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, Klein MD. Automated robot-assisted surgical skill evaluation: predictive analytics approach. Int J Med Robot. 2018 Feb; 14(1):e1850.
16 Zia A, Essa I. Automated surgical skill assessment in RMIS training. Int J CARS. 2018 May; 13(5):731–9.
17 Jarc AM, Curet MJ. Viewpoint matters: objective performance metrics for surgeon endoscope control during robot-assisted surgery. Surg Endosc. 2017 Mar;31(3):1192–202.
18 Hung AJ, Chen J, Jarc A, Hatcher D, Djaladat H, Gill IS. Development and validation of objective performance metrics for robot-assisted radical prostatectomy: a pilot study. J Urol. 2018 Jan;199(1):296–304.
19 Lyman WB, Passeri M, Murphy K, Siddiqui IA, Khan AS, Lannitti DA, et al. Novel objective approach to evaluate novice robotic surgeons using a combination of kinematics and stepwise cumulative sum analyses. J Am Coll Surg. 2018;227(4):S223–4.
20 Padoy N. Machine and deep learning for workflow recognition during surgery. Minim Invasive Ther Allied Technol. 2019 Apr; 28(2):82–90.
21 Sarikaya D, Jannin P. Towards generalizable surgical activity recognition using spatial temporal graph convolutional networks. arXiv preprint. 2020 Aug. arXiv:2001.03728.
22 DiPietro R, Lea C, Malpani A, Ahmidi N, Vedula SS, Lee GI, et al. Recognizing surgical activities with recurrent neural networks. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. 19th International Conference, Proceedings. Cham: Springer; 2016. pp. 551–8.
23 van Amsterdam B, Nakawala H, Momi ED, Stoyanov D. Weakly supervised recognition of surgical gestures. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE. 2019. p. 9565–71.

24 Zia A, Zhang C, Xiong X, Jarc AM. Temporal clustering of surgical activities in robot-assisted surgery. Int J CARS. 2017 Jul;12(7):1171–8.
25 Zia A, Hung A, Essa I, Jarc A. Surgical activity recognition in robot-assisted radical prostatectomy using deep learning. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Lecture Notes in Computer Science. Vol 11073. Cham: Springer; 2018. p. 273–280.
26 Zia A, Guo L, Zhou L, Essa I, Jarc A. Novel evaluation of surgical activity recognition models using task-based efficiency metrics. Int J CARS. 2019 Dec;14(12):2155–63.
27 Maier-Hein L, Vedula S, Speidel S, Navab N, Kikinis R, Park A, et al. Surgical data science: enabling next-generation surgery. arXiv preprint. 2017. arXiv:1701.06482.
28 Chen J, Cheng N, Cacciamani G, Oh P, Lin-Brande M, Remulla D, et al. Objective assessment of robotic surgical technical skill: a systematic review. J Urol. 2019 Mar;201(3):461–9.
29 Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18(17):1–5.
30 Laurikkala J. Improving identification of difficult small classes by balancing class distribution. In: Quaglini S, Barahona P, Andreassen S, editors. Artificial Intelligence in Medicine. Berlin, Heidelberg: Springer; 2001. p. 63–6.
31 Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. 2004 Jun;6(1):20–9.
32 James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Vol 112. New York: Springer; 2013.
33 Abdi H. Bonferroni test. In: Salkind NJ. Encyclopedia of measurement and statistics. Vol 3. Sage Publications; 2007. p. 103–7.
34 Guan D, Yuan W, Lee YK, Najeebullah K, Rasel MK. A review of ensemble learning based feature selection. IETE Tech Rev. 2014;31(3):190–8.
35 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
36 Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition, Istanbul. IEEE. 2010. p. 3121–24.
37 Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975 Oct;405(2):442–51.
38 Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020 Jan;21(1):6.
39 Ghani KR, Miller DC, Linsell S, Brachulis A, Lane B, Sarle R, et al.; Michigan Urological Surgery Improvement Collaborative. Measuring to improve: peer and crowd-sourced assessments of technical skill with robot-assisted radical prostatectomy. Eur Urol. 2016 Apr;69(4):547–50.
40 Peabody JO, Miller DC, Linsell S, Lendvay T, Comstock B, Lane B, et al. Wisdom of the crowds: use of crowdsourcing to assess surgical skill of robot-assisted radical prostatectomy in a statewide surgical collaborative. Eur Urol Suppl. 2015;14(2):e192–e192a.
41 Powers MK, Boonjindasup A, Pinsky M, Dorsey P, Maddox M, Su LM, 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 Apr;30(4):447–52.
42 Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol. 2012 Jan;187(1):247–52.
43 Grantcharov TP, Reznick RK. Teaching procedural skills. BMJ. 2008 May;336(7653):1129–31.
44 Gao Y, Vedula SS, Reiley CE, Ahmidi N, Varadarajan B, Lin HC, et al. JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling. In: MICCAI Workshop: M2CAI. 2014;3:3.
45 Liu M, Purohit S, Mazanetz J, Allen W, Kreaden US, Curet M. Assessment of Robotic Console Skills (ARCS): construct validity of a novel global rating scale for technical skills in robotically assisted surgery. Surg Endosc. 2018 Jan;32(1):526–35.