HHS Public Access

Author manuscript

JAMA Surg. Author manuscript; available in PMC 2022 May 09.

Published in final edited form as:

JAMA Surg. 2018 August 01; 153(8): 770–771. doi:10.1001/jamasurg.2018.1512.

Automated Performance Metrics and Machine Learning Algorithms to Measure Surgeon Performance and Anticipate Clinical Outcomes in Robotic Surgery

Andrew J. Hung, MD, **Jian Chen, MD**, **Inderbir S. Gill, MD**

Center for Robotic Simulation and Education, Catherine and Joseph Aresty Department of Urology, USC Institute of Urology, University of Southern California, Los Angeles.

What Is the Innovation?

Mounting research confirms that surgeon performance is directly associated with patient outcomes.[1] The current criterion standard for evaluating surgeons is peer review, either during surgery or retrospectively via video footage. Expert review is also used to evaluate performance on robotic surgery. Yet systems data captured directly from the robot provide a novel opportunity to more accurately and objectively measure surgeon performance. A method using data from the robot could increase accuracy and decrease reliance on expert evaluators. We used a novel da Vinci Systems recording device (dVLogger; Intuitive Surgical, Inc) to collect automated performance metrics (APMs) (instrument and endoscopic camera motion tracking and events data, such as energy usage) during live robotic surgery.[2] We used machine learning (ML) algorithms—now commonplace outside of medicine—to process these large volumes of automatically collected data (Figure). Machine learning, a form of artificial intelligence, relies on computer algorithms and large volumes of data to "learn" and recognize broad patterns that are often imperceptible to human reviewers. With this process, we can now objectively measure surgeon performance and anticipate patient outcomes; in the near future, we will be able to personalize surgeon training.

What Are the Key Advantages Over Existing Approaches?

Currently, the most practical way to estimate surgical performance is by the surgeon's previous case volumes: a surgeon with a high case volume is likely to perform better than one with a low case volume. Peer assessment of video recordings can also estimate surgeon expertise and possibly anticipate outcomes. This method is subject to interobserver variability because expert surgeons often disagree about what constitutes good surgery.[3] Combining procedure-specific APMs with ML algorithms can produce a truly objective assessment of surgeon performance. Automated assessment with minimal human processing (and thus with minimal bias introduced) can provide valuable feedback,

Corresponding Author: Andrew J. Hung, MD, Center for Robotic Simulation and Education, Catherine and Joseph Aresty Department of Urology, USC Institute of Urology, 1441 Eastlake Ave, Ste 7416, Los Angeles, CA 90089 (andrew.hung@med.usc.edu).

Additional Contributions: Yan Liu PhD, Zhengping Che, BE, and Tanachat Nilanon, MS (USC Machine Learning Center), trained the machine learning model. Anthony Jarc, PhD (Intuitive Surgical, Inc), supported the dVLogger devices that recorded the da Vinci Systems data and was compensated.

both to individual surgeons and to credentialing and licensure committees. Assessment tools that use automation and self-learning computer algorithms can provide a sustainable method for large-scale evaluation of surgeons. In 2016, 563 000 robotic operations were performed in the United States.[4] Although it is not necessary to evaluate every surgical procedure, periodic reevaluations would ensure that surgeons maintain proficiency. Automated performance metrics may provide a more comprehensive and objective picture of a surgeon's skills than expert evaluators. Subtle fluctuations in performance may escape detection until after an adverse clinical outcome occurs.

Automated performance metrics evaluate the actual robotic surgical performance. Although easy to collect, these vast volumes of data require specialized methods for efficient processing. Machine learning is a natural fit for analyzing large data sets, with several advantages over conventional statistics.[5] Conventional analyses require the a priori selection of a model most suitable for the study data set. In addition, only significant or theoretically relevant variables based on previous experience are included for analysis. In contrast, ML is not built on a prestructured model; rather, the data shape the model by detecting underlying patterns. The more variables (input) used to train the model, the more accurate the ultimate model will be. Machine learning algorithms require that training material be tagged with corresponding labels. With APMs as the training material, the ideal label is *patient outcomes*. This tagging ensures that all surgeon evaluations through ML retain clinical relevance.

How Will This Affect Clinical Care?

Nuanced surgeon performance data, harnessed correctly, can personalize training in precise and effectual ways. Guided by surgical educators, ML models can identify performance qualities not necessarily evident to experienced trainers. For example, we have already identified 2 specific factors. First, conventional wisdom accepts that bimanual dexterity (ie, balanced use of both hands) is an ideal surgical trait. To the contrary, our pilot APM data found that expert surgeons use their dominant hand more than novices.[2] Second, by ranking metrics adopted by ML algorithms for relevance, we can identify specific, desirable performance qualities. For example, we identified metrics related to camera manipulation that strongly correlate with surgeon expertise and good outcomes. In reality, these camera manipulation metrics may represent a sensitive aggregate measure of surgeon performance rather than a specific technical skill to develop in surgeons. As always, associations do not automatically represent cause and effect and must be taken judiciously.

Improved personalized assessment and training of surgeons should improve patient safety on a large scale. Automated performance metrics can provide an additional layer of assessment beyond peer review for robotic surgery credentialing or licensure. Such evaluation would provide an additional safety check for surgeons before they may operate on patients.[6]

What Evidence Supports the Benefits of the Innovation?

Automated performance metrics are already validated in the laboratory.[7] To our knowledge, our pilot data, captured during live surgery, are the first to correlate APMs to patient

outcomes.[8] Calculations were made using the scikitlearn package in Python, version 0.19.1 (Python Software Foundation). Using a random forest 50-tree ML model based on APMs alone, we can calculate with 87.2% accuracy whether the hospital stay of a patient undergoing robot-assisted radical prostatectomy will be 2 or fewer or more than 2 days (Figure). Inclusion of patient characteristics (age, body mass index, prostate-specific antigen level, and prostate volume) increased accuracy to 88.5%. In addition, through the ML model, we can identify top APMs associated with clinical outcomes. As longer-term clinical data mature in our series, we will be able to use labels such as *oncologic outcomes* and *functional outcomes*.

What Barriers Prevent Implementation More Broadly?

Training ML algorithms requires expertise and collaboration between computer scientists and surgical educators. Previous efforts primarily involved engineers in the laboratory setting lacked clinical relevance. The balance between too little and too much human handling for ML algorithms is delicate. With too little, the output lacks practical use; too much handling may introduce bias or misinformation. Existing ML algorithms outside of medicine should be adapted to clinical needs. We are leading an ongoing 5-year multi-institutional study that involves several high-volume international groups and aims to deliver a robustly trained model with a multi-institutional, multisurgeon automated data set with accompanying clinical outcomes.

When Will This Innovation Likely Be Applied Routinely?

With continued development of robotic surgery and ML algorithms—on parallel paths of innovation—we anticipate processed automated data will accurately measure surgeon expertise and anticipate surgical outcomes within 3 to 5 years.

Funding/Support:

This work was funded by USC. Intuitive Surgical, Inc, provided dVLogger recorders to collect the data.

Role of the Funder/Sponsor:

USC had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Conflict of Interest Disclosures:

Dr Hung reported serving as a consultant for Ethicon, Inc and receives departmental clinical research funding from Intuitive Surgical, Inc. No other disclosures were reported.

REFERENCES

1. Birkmeyer JD, Finks JF, O'Reilly A, et al. ; Michigan Bariatric Surgery Collaborative. Surgical skill and complication rates after bariatric surgery. N Engl J Med. 2013;369(15):1434–1442. doi:10.1056/NEJMsa1300625 [PubMed: 24106936]
2. 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;199(1): 296–304. doi:10.1016/j.juro.2017.07.081 [PubMed: 28765067]

3. Lendvay TS, White L, Kowalewski T. Crowdsourcing to assess surgical skill. JAMA Surg. 2015;150(11):1086–1087. doi:10.1001/jamasurg.2015.2405 [PubMed: 26421369]

4. Intuitive Surgical, Inc. 2016 Annual Report. http://www.annualreports.com/HostedData/AnnualReports/PDF/NASDAQ_ISRG_2016.pdf. Accessed May 16, 2018.

5. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–2410. doi:10.1001/jama.2016.17216 [PubMed: 27898976]

6. Tam V, Zeh HJ III, Hogg ME. Incorporating metrics of surgical proficiency into credentialing and privileging pathways. JAMA Surg. 2017;152(5): 494–495. doi:10.1001/jamasurg.2017.0025 [PubMed: 28273289]

7. Kumar R, Jog A, Vagvolgyi B, et al. Objective measures for longitudinal assessment of robotic surgery training. J Thorac Cardiovasc Surg. 2012;143 (3):528–534. doi:10.1016/j.jtcvs.2011.11.002 [PubMed: 22172215]

8. Hung AJ, Chen J, Che Z, et al. Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. J Endourol. 2018;32(5):438–444. doi:10.1089/end.2018.0035 [PubMed: 29448809]

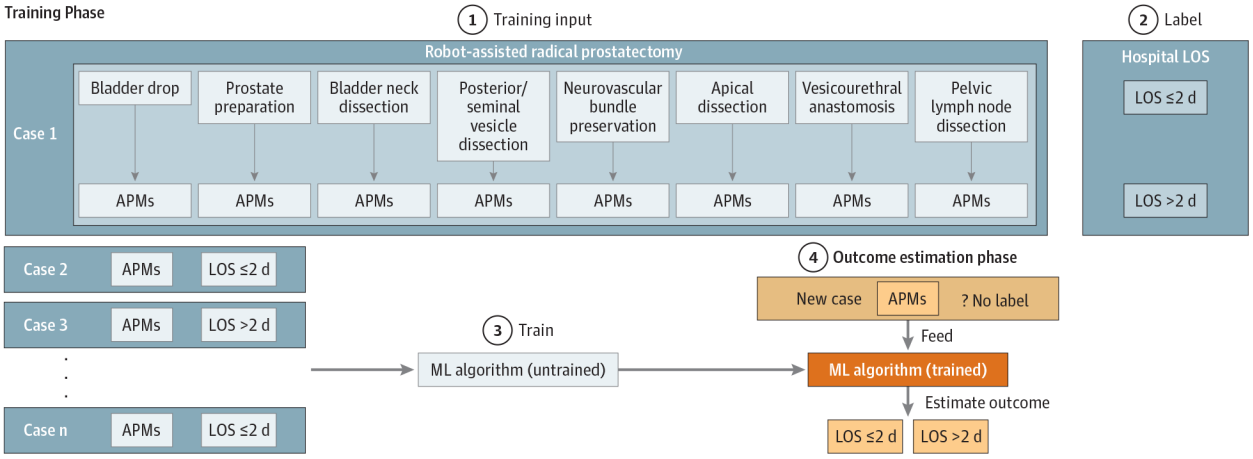

Figure.

Training a Machine Learning (ML) Algorithm to Estimate Perioperative Clinical Outcome With Automated Performance Metrics (APMs)

(1) A set of 25 APMs is captured from each of 8 distinct surgical steps during robot-assisted radical prostatectomy. (2) Cases are labeled as LOS ≤2 d or LOS >2 d based on hospital length of stay (LOS) after surgery. (3) A random forest 50-tree ML algorithm is trained with APMs and the clinical outcome label. (4) The trained ML algorithm can then estimate clinical outcomes (LOS ≤2 d or LOS >2 d) of a new case by using APMs only.