Received: 1 March 2021 | Revised: 29 March 2021 | Accepted: 2 April 2021

DOI: 10.1002/jso.26496

REVIEW ARTICLE

Journal of SURGICAL ONCOLOGY WILEY

Surgical data science and artificial intelligence for surgical education

Thomas M. Ward MD[1] | Pietro Mascagni MD[2,3,4] | Amin Madani MD, PhD[5] | Nicolas Padoy PhD[2,4] | Silvana Perretta MD[4,6] | Daniel A. Hashimoto MD, MS[1]

[1]Department of Surgery, Surgical AI & Innovation Laboratory, Massachusetts General Hospital, Boston, Massachusetts

[2]ICube, University of Strasbourg, CNRS, France

[3]Fondazione Policlinico A. Gemelli IRCCS, Rome, Italy

[4]IHU Strasbourg, Strasbourg, France

[5]Department of Surgery, University Health Network, Toronto, Canada

[6]IRCAD, Strasbourg, France

Correspondence
Daniel A. Hashimoto, MD MS, Surgical AI & Innovation Laboratory, Department of Surgery, Massachusetts General Hospital, 15 Parkman St, WAC460, Boston, MA 02114, USA.
Email: dahashimoto@mgh.harvard.edu

Abstract

Surgical data science (SDS) aims to improve the quality of interventional healthcare and its value through the capture, organization, analysis, and modeling of procedural data. As data capture has increased and artificial intelligence (AI) has advanced, SDS can help to unlock augmented and automated coaching, feedback, assessment, and decision support in surgery. We review major concepts in SDS and AI as applied to surgical education and surgical oncology.

KEYWORDS

artificial intelligence, computer vision systems, data science, deep learning, surgical education

1 | INTRODUCTION

Surgical data science (SDS) aims to improve the quality of interventional healthcare and its value through the capture, organization, analysis, and modelling of procedural data. Although this certainly includes surgery, the tent of SDS also includes other procedural fields such as interventional radiology, pulmonology, and gastroenterology.[1] Building on the evolution of innovative technologies in these fields and the foundation of data analysis established by the quality improvement and health services fields, SDS incorporates a range of inputs from traditional registry and claims data, device data, and patient/surgeon-specific data to yield insights into procedural care (Figure 1).

Over time, clinicians have evolved from relying only on intuition and personal experience to having data-driven intuition and collective experience as the result of technology, innovation, and evidence-based medicine. The trick now is striking the correct balance from population-level data and clinician experience to deliver individualized care (Figure 2) as there is no area of medicine that is currently "data only." As clinicians, we are expected to consider the best available evidence or data for an individual patient and then call upon our experience to ultimately make a decision in the best interest of the specific patient.

Artificial intelligence (AI) has also recently grown in popularity and interest within surgery. Although SDS and AI in surgery are not one and the same, SDS can utilize techniques from AI to facilitate improvements in the delivery of surgical care, whether through direct patient contact via diagnostic and therapeutic intervention or through clinicians who can benefit from data-enabled insights into their own performance. Other published studies have conducted reviews of the diagnostic and therapeutic potential of SDS and AI[2–5]; we focus our review on applications to surgical education, including decision support and coaching, feedback, and performance assessment. We briefly review concepts in AI in surgery, their current applications in research (emphasizing possible impacts for surgical education), and the anticipated future directions of the field.

1.1 | AI in surgery

The field of AI studies how to make computers function intelligently to understand, process, and act in the world. The media portrays a

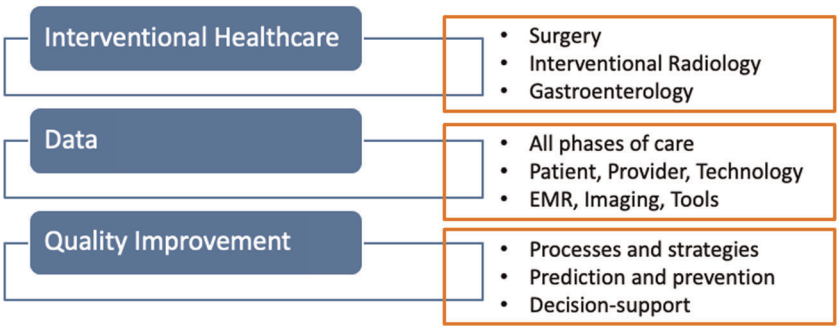

FIGURE 1 Surgical data science (SDS) covers interventional healthcare and draws on multiple sources of data to yield quality improvement. Courtesy of SAIIL. Reproduced with permission [Color figure can be viewed at wileyonlinelibrary.com]

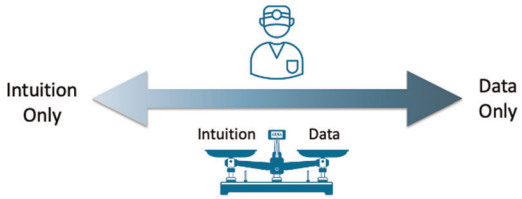

FIGURE 2 Modern clinicians have varying levels of intuition or experience and data depending on the clinical situation and must balance the two to deliver care to patients [Color figure can be viewed at wileyonlinelibrary.com]

yet-to-be-achieved level of AI, Artificial General Intelligence, where robots and computers function at human-like levels.[6] Realistically, current AI functions in a narrow manner, where computers can successfully perform a few select tasks. Today, narrow AI is ubiquitous: it recommends movies to consumers, drives autonomous vehicles, and even processes letters for postal services.

Although the term AI was not formally coined until 1956 at the Dartmouth Summer Research Project, influences on the field date back decades (if not centuries).[7] Despite lofty initial promises, AI as a field experienced a "boom and bust" trajectory over the next four decades, including two "AI Winters" during which excitement and funding for the field dropped significantly.[8] The most recent "AI Winter" has experienced a thaw over the past twenty-five years, driven by gradual advances in three combined factors: data availability, computational power, and novel techniques.

Data forms the foundation upon which computers learn. Now, more than ever, we generate "big data." Before 2003, humanity created 5 exabytes ($5*10^{18}$) of data -- an amount now generated every *two* days.[9] Advances in deep learning, which started in the 1990s, created algorithms with multiple processing layers that could use this tremendous amount of data to discover complex and hidden patterns.[10] However, we could not use these algorithms efficiently until recently with the use of modern high-powered computer chips called Graphical Processing Units (GPUs) that can process the mountains of data through the multiple layers of deep learning.[11]

Machine learning (ML) is a field of study within AI that tries to teach computers and machines how to learn. Learning is a fundamental process where an agent (be it human or computer), uses information from observations to improve its performance on future tasks. Machine learning is often categorized into four different categories described by the feedback with which the machine tries to learn: supervised, semi-supervised, unsupervised, and reinforcement.[12] In supervised learning, the machine algorithm receives inputs in the form of labeled (annotated) data. It learns the ability to take future unlabeled data and correctly determine the labels. For example, a machine could be fed x-rays labeled with pneumonia and could subsequently attempt to identify pneumonia in unlabeled x-rays.

In unsupervised learning, machines recognize patterns within their input data. As an example, if given a wide sample of unlabeled chest radiographs, the machine will group the films into many categories. Upon review, these categories may include films that have "no pathology" or "pneumonia," but they may even uncover previously unknown clusters (e.g., subtle signs of volume overload or early malignancies). Semi-supervised learning lies in-between the two above extremes, training on partially labeled data.

Lastly, reinforcement learning occurs via trial and error while attempting to achieve a specific objective. The AI either receives a positive reward when it succeeds or punishment when it fails. In reinforcement learning, the machine is not told exactly what it did wrong, but rather, just that it did something wrong. It then refines its actions over repeated iterations, keeping those that led to success and avoiding those that led to failure. Although reinforcement learning seems to be an indirect and therefore inefficient way to train an AI, it can produce incredible results. Through reinforcement learning, a computer algorithm with no other precedent knowledge of the game of Go was able to easily beat the world champion.[13]

Classical ML methods use structures such as trees to represent data and require experts in the field to "hand-craft" parts of the process (e.g., select features most likely to represent the phenomenon of interest) to work with each domain's data. Neural networks take a different approach, where inferences are made from the data itself without requiring manual feature selection. Neural networks have proven to be highly adaptable and generalizable to automatically select the features that are most likely to yield results for a given phenomenon of interest; thus, they can work for a multitude of problems.[7]

Similar to a human brain's internal structure, a neural network is composed of tens to hundreds, and even thousands, of computer-represented "neurons" that are either on or off. A deep-learning network is a stack of three or more neuronal layers, with each layer specializing at a certain task, creating more specific output through each

progression of the layers.[10] To provide a simplified example, consider a deep learning network with multiple layers for recognizing geometric shapes. The first layer may decide if an object has straight versus rounded edges. Having noted straight edges, the next layer determines the number of edges (in this case four). The ensuing layer detects the four edges are the same length, and the final layer detects each edge is at a right-angle to the other. Taken all together, the network finally outputs the object's classification as a square (Figure 3).

The beauty of deep learning lies in the summation of each layer, which performs a small simple function, into a complex overall output. As it is composed of small functions, different layers can be combined to produce different outputs, and even small modifications of just the last few layers allow for easy reapplication of a model that performs one task (such as diagnosing chest radiographs) to a model that can accurately perform another (such as diagnosing retinal images). This technique is known as transfer learning, and the prior example in fact has been published in the literature.[14,15] The use of deep learning has particularly revolutionized the fields of image recognition and language processing.

Although classical machine learning, neural networks, and deep learning are not the only approaches to AI available to researchers and developers, they are some of the most popular in use in surgery.[4] Furthermore, each of these techniques can be utilized within applicative fields such as natural language processing and computer vision, as well as strengthen the impact of other fields related to real-world decision making. By processing and inputting real-world data into a form in which computers can reason or better "understand" the impact of actions, in reality, advances in the fields are expected to significantly impact surgery and many other application domains.

Natural language processing (NLP), the comprehension of written human language by computers, has also undergone a revolution due to neural networks. Before neural networks, NLP was limited to the categorization of language in N-gram models, which is the probability of an ensuing word as a result of the preceding words.[12] Language, however, is highly composable and does not lend itself to the prediction from the preceding words alone that N-gram models generate. The composable nature of neural networks nicely mirrors that of language, which allows for extremely accurate NLP.[16] Now that computers can reliably model and understand language, researchers have begun to deploy NLP in multiple medical applications, particularly those related to the electronic medical record.

Computer vision (CV) is the process of training a computer to see and understand images. After relative stagnation for decades, Krizhevsky et al.[17] revolutionized the field in 2012, obtaining human-like object classification accuracy. Their algorithm succeeded through the utilization of a particular neural network structure called convolutional neural networks (CNNs). CNNs work in a similar fashion to a human's visual cortex. Instead of needing to process every pixel of information, they allow for classification of an image's parts into the key components necessary for recognition, such as shape, texture, and color. As the network only needs to train on these small components, it learns more efficiently and quickly while attempting to minimize the dangers of overfitting. Overfitting occurs when a statistical model conforms too closely to a selection of data points and, therefore, performs poorly when applied to other data sets. As CNNs learn to visually recognize objects through the key components alone, they can recognize a car, for example, even if it is a different brand or color. The success of CNNs has led to numerous deployments in visual aspects of medicine, from image-based diagnostics to real-time surgical video analysis. It has also formed the basis for much of the recent success and the growing interest in AI in surgery.

1.2 | AI for surgical performance augmentation and education

Data over the last three decades have shown alarmingly high rates of preventable adverse events amongst hospitalized surgical patients.

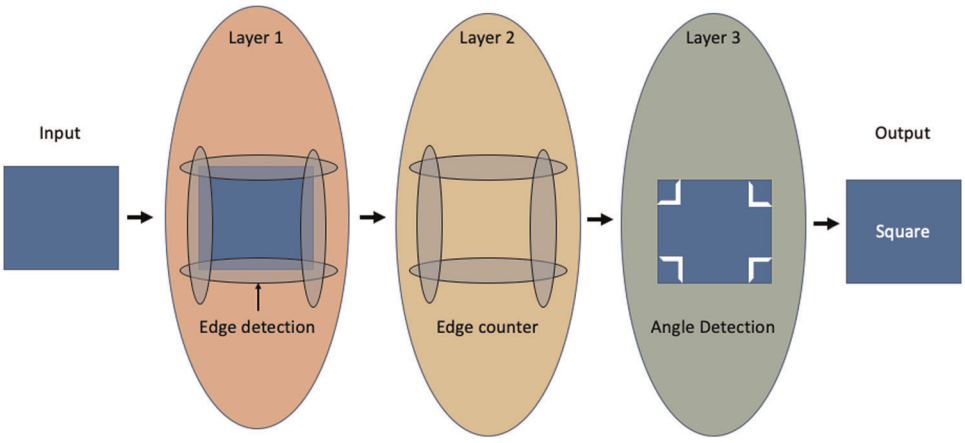

FIGURE 3 Simplified schematic of a neural network that identifies shapes. Each layer derives a piece of information that can ultimately result in the identification of a shape such as a square. More complex versions of this basic architecture have achieved tasks such as identification of surgical instruments, anatomy, and operative steps. Courtesy of SAIIL. Reproduced with permission [Color figure can be viewed at wileyonlinelibrary.com]

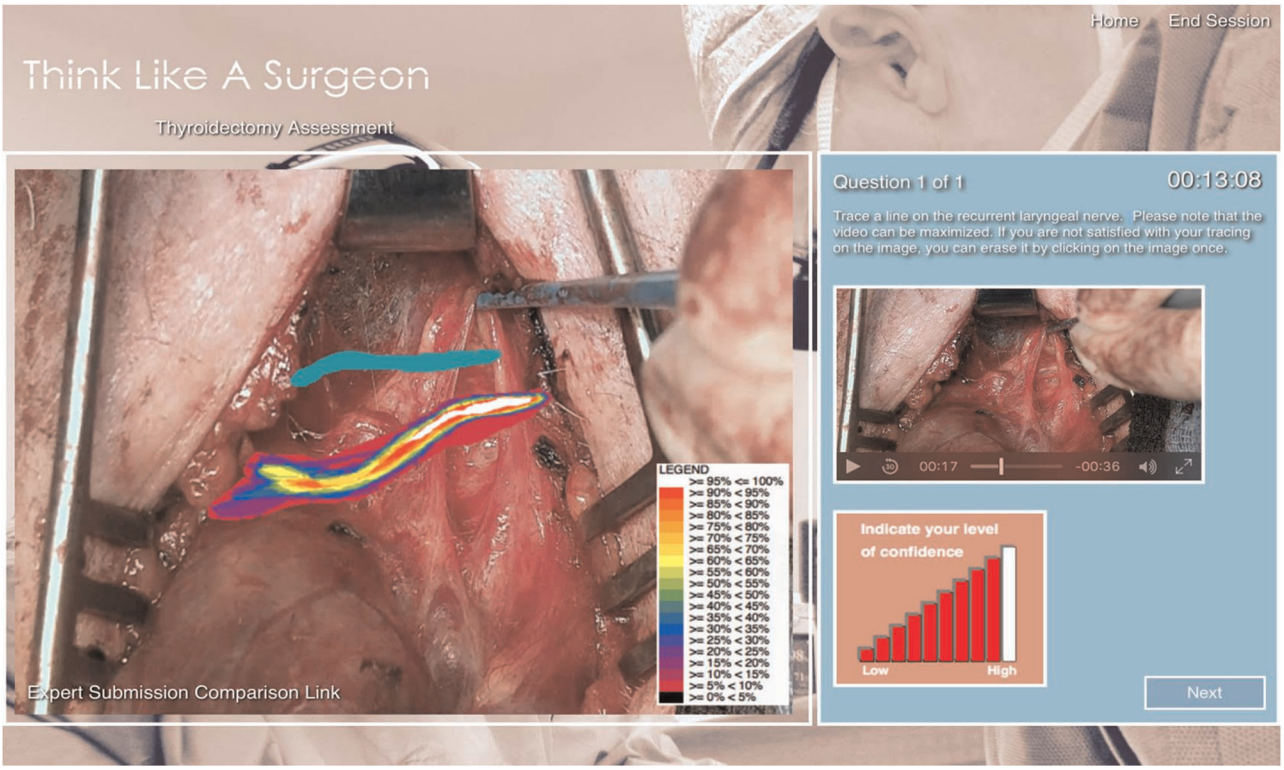

FIGURE 4 Screenshot of the think like a surgeon platform demonstrating annotation of recurrent laryngeal nerve by multiple experienced surgeons to generate a "heat map" of areas where the nerve is perceived to be. Courtesy of Amin Madani. Reproduced with permission [Color figure can be viewed at wileyonlinelibrary.com]

Within these reports, which include a wide range of demographics, geographic locations, and surgical subspecialties, root-cause analyses tend to trace most of the errors back to events that occurred at the time of surgery.[18–21] Specifically, adverse events tend to occur as an error in judgment or decision-making that led to behaviors or actions that contributed to the outcome.[22] Most surgeons would agree that the skills that contribute the greatest to the development of an elite surgeon are their cognitive skills as opposed to psychomotor skills. This is also supported by the body of literature in surgical education emphasizing the importance of intraoperative judgment and decision-making as a dominant determinant of surgical performance and, ultimately, surgical outcome.[22] For instance, errors in human visual pattern recognition can lead to misinterpretation of surgical anatomy that leads to injury of a critical structure, such as the bile duct during laparoscopic cholecystectomy.[23]

It is, therefore, unsurprising that with the advances in the digitalization of the surgical field and the ability to collect large sets of data from the operating room (OR) (e.g., images, videos), innovators and researchers have turned to machine learning as a potential tool to augment a surgical team's performance. The field of computer vision has made significant strides over the last decade, and new machine learning methodologies such as deep learning have provided the means to develop algorithms that can perform advanced human-level perceptual functions, such as object recognition and tracking within a video, and scene recognition. Given that most errors stem from advanced cognitive skills, there is tremendous potential in developing algorithms that are able to analyze surgical data and augment our mental model to improve surgical decision-making.

Although computer vision has shown promise in various non-surgical fields of medicine (e.g., cancer diagnosis from mammograms), there are several challenges to consider before applying computer vision to surgery. Firstly, anatomical structures in the field are almost never well-demarcated and often hidden under layers of fatty and fibrous tissues, making it difficult to train a model for intraoperative guidance and navigation. To compound this problem, unlike images from diagnostic radiology or fundoscopy, most surgical videos have a tremendous amount of variability in terms of background noise, image and video quality, and other objects in the field. Secondly, there is a significant variation amongst experts with respect to their advanced cognitive behaviors. For instance, most experts will not agree on the exact location where to dissect, or the exact location of an anatomical plane, or the best possible instrument to use in any given instance. Therefore, establishing a gold-standard reference ("ground truth") on which to train an algorithm and evaluate its performance is a major obstacle.

To overcome some of these challenges, Madani et al.[24] proposed the visual concordance test (VCT) as a novel methodology to establish expert consensus within a surgical field. In this process, surgeons make annotations on frames extracted from a surgical video, while watching the video itself for reference. These annotations can

be obtained from a panel of experienced surgeons and compiled to create a "heat map" that demonstrates the level of agreement amongst these surgeons (Figure 4). Despite the fact that most experts will not annotate the exact same set of pixels, the area of convergence of these annotations is considered to be an agreement amongst the expert panel, and these pixels are subsequently used to train the AI model (e.g., the area the panel agrees is the best location to dissect in that particular scene).

One application that leverages the lessons learned from the mapping of experts' mental models and surgical decision-making using VCT is the development of GoNoGoNet and CholeNet.[25] These models were developed to automatically detect and outline safe areas of dissection ("Go zone"), dangerous areas of dissection ("No-Go zone"), and other anatomical structures during laparoscopic cholecystectomy. In this study, a data set of 290 laparoscopic cholecystectomy videos drawn from 136 institutions in 37 countries was used to train these models with over 90% pixel accuracy and good spatial overlap compared to ground truths (Figure 5A,B). Real-time overlay of Go and No-Go zones could provide feedback and guidance to surgeons who wish to learn new operations, seek to improve their performance, or find themselves in particularly difficult operations.

In light of the encouraging results from GoNoGoNet, the applications for surgical oncology can be potentially transformational. Performing an adequate oncologic resection while minimizing perioperative morbidity is the cornerstone of most cancer operations. For example, deviation from the ideal dissection plane can either lead to an oncologically inadequate resection or an increased risk of complications due to injury to surrounding structures. Several groups are currently working on developing models to provide real-time guidance on the ideal dissection plane during cancer operations to minimize early perioperative outcomes and improve long-term oncologic outcomes.

Surgical decision-making is not always a cognitive behavior that relates to a specific location in the surgical field (e.g., where to dissect). Often it occurs at a higher level in relation to the tactical approach of the operation. For this reason, AI-based automated scene recognition and assessment could also be leveraged to assist in critical decision points of procedures, especially when significant operator variability exists. For instance, during laparoscopic cholecystectomy, it is not only important to keep the dissection in a safe plane that minimizes the risk of a major bile duct injury, but it is also important not to divide any cystic structures until a critical view of safety (CVS) has been achieved. It is also important to consider various bailout procedures if a hostile environment is encountered and a CVS cannot be safely achieved, such as a subtotal cholecystectomy.[26,27] Given that the determination of a CVS has been found to be highly operator dependent,[28,29] a model that provides decision-support to surgeons in real-time as to what is the most optimal strategy could be highly advantageous. Mascagni et al. recently published their results on DeepCVS, which is a two-stage model to segment (i.e., delineate an object along its boundaries) hepatocystic anatomy and predict whether or not each of the three elements of the CVS has been achieved (Figure 6).[30] The model had a mean average precision greater than 70%, suggesting that such a technology could potentially augment intraoperative judgment for difficult situations.

1.3 | Automated phase and instrument recognition

Today, instead of using the rich information from intraoperative events, most of the surgery is distilled into a one-page operative report. These operative reports fail to detail almost a third of intraoperative complications.[31] They also fail to capture how the operation proceeded, as in, how well did the surgeon perform the operation. We know intraoperative performance matters: across a group of bariatric surgeons, those in the top quartile of surgical skill had lower rates of reoperation, readmission, ED visits, surgical complications, and medical complications.[32] This has been replicated in colorectal surgery for transanal total mesorectal excision with Curtis et al demonstrating that surgeons in the upper quartile of technical skill had better outcomes as measured by integrity of the mesorectal dissection plane and morbidity.[33] Although surgical coaching programs are growing in popularity and could provide the means through which surgical performance could be improved (especially in practicing surgeons), video review remains a tedious exercise.[34–36] SDS could assist in the analysis of surgical skills through automated methods to segment, annotate, and assess operative video.

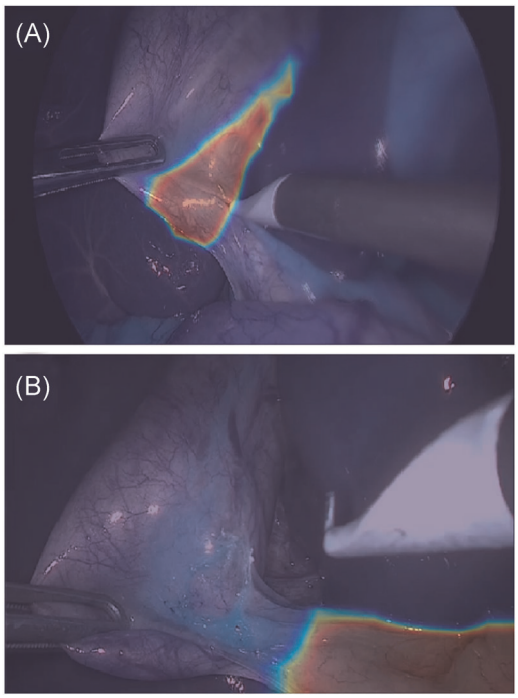

FIGURE 5 (A) Image demonstrating a "Go" zone or safe area of dissection identified by GoNoGoNet and (B) a "NoGo" zone or unsafe area of dissection. Courtesy of Amin Madani. Reproduced with permission [Color figure can be viewed at wileyonlinelibrary.com]

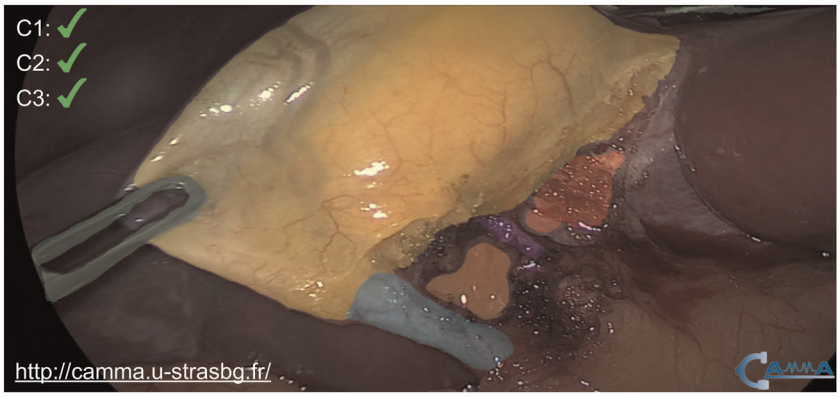

FIGURE 6 Example of DeepCVS semantic segmentation of the gallbladder (yellow), the cystic duct (cyan), the cystic artery (blue), the dissected hepatocystic triangle (orange), the exposed cystic plate (red), and surgical instrument (light green) to improve the performance and interpretability of the automatic assessment of critical view of safety criteria (C1: 2 tubular structures connected to the gallbladder; C2: a well dissected hepatocystic triangle; C3: the lower part of the gallbladder is dissected to expose the cystic plate). Courtesy of CAMMA. Reproduced with permission [Color figure can be viewed at wileyonlinelibrary.com]

As anecdotally known by surgical educators and formalized in the Zwisch model for teaching and assessment in the OR, understanding surgical instrument usage and the sequence of steps needed to successfully complete a surgical procedure is an early stage of surgical training.[37] Subsequently, awareness and anticipation of procedures' workflows are gained before ultimately having trainees actively participating in surgical interventions.[38]

Inspired by this intuition in procedural skill development, the SDS community has firstly focused on developing computer vision systems for automatic phase recognition and instrument detection, fundamental elements of surgical workflows,[39] with the aim to then provide context-aware assistance in the OR.[39,40] Starting from fairly standardized procedures such as cholecystectomy and hysterectomy, early algorithms made use of hand-crafted signals such as surgical instrument usage and time dependencies to model and visualize surgical workflow using classical machine learning techniques such as dynamic time warping and hidden Markov models.[41,42] More recently, breakthroughs in deep neural networks have boosted computer vision performance and revived the field of surgical workflow analysis.[43]

In 2016, Twinanda et al trained a convolutional neural network on 80 publicly released laparoscopic cholecystectomy videos (Cholec80) to detect surgical instruments and recognize seven phases of the procedure in a multitasking manner.[44] The resulting model, EndoNet, demonstrated 92% accuracy in the classification of surgical phases when used for posthoc analysis, and 86% accuracy when used for real-time inference.[44] Deep learning models have since been trained to accurately recognize phases across a range of procedures, including bariatric,[45] colorectal,[46] ophthalmic,[47] and intraluminal interventions.[48]

These models were trained using supervised learning with annotations of video that contain the phase or the surgical instrument seen in a given video frame. The time-consuming and tedious process of manually labeling images has so far limited the informativeness and size of annotated surgical datasets, hampering the development of AI models capable of performing more demanding surgical tasks and limiting the generalization of such models across centers and procedures. The SDS community has been working to overcome this limitation by devising models that learn using less (i.e., weakly-supervised) or no (i.e., self-supervised) manual annotations. For instance, Yu et al. used a small data set of annotated videos to train a "teacher" model to automatically label a larger set of videos for training a lighter "student" model capable of real-time inference, utilizing state-of-the-art models for surgical instrument detection, localization, and tracking that were trained using images annotated only with binary instruments presence information (i.e., a given tool is present or not in an image).[49–51]

Automated workflow analysis could serve surgical education in a multitude of ways, most of which have yet to be explored. Potential pedagogic uses of phase recognition and instrument detection models can be categorized based on the timing of the analysis—either post hoc for feedback or real-time for guidance and decision-support.

Post hoc analysis of surgical workflows has been proposed to facilitate video-based assessment (VBA). This valuable approach for the evaluation of performance and quality improvement is currently limited by the time-consuming process of collecting, manually reviewing, and editing long surgical videos.[36] Automated workflow inference could be used to make VBA more efficient, standardized, and scalable. Phases and instrument usage information could be used to synchronize videos of the same surgical procedure, allowing for smart indexing and efficient querying of large databases of surgical videos. As recently shown by a Japanese group analyzing gastrectomies,[52] surgical instrument usage patterns could be plotted to efficiently screen for cases and scenes likely to show unexpected events so as to prioritize their VBA. The Surgical AI and Innovation Laboratory at Massachusetts General Hospital have described the concept of the "surgical fingerprint" wherein phase recognition

algorithms can assess the video of interest's workflow against that of a pre-existing database to determine whether the video is following an expected operative course. Time points with deviations from an expected operative course could signal areas where complications or unexpected events might occur.[45,48]

To demonstrate the potential value of such workflow information in rapidly analyzing operative video, a computer vision platform called EndoDigest used the predictions of phase recognition and instrument detection models to accurately detect the time of the cystic duct division in laparoscopic cholecystectomy videos and automatically provided short videos effectively documenting the critical view of safety in 91% of cases (Figure 7).[53] These and other similarly "digested" surgical videos could be used for auditing performance and/or efficiently rehearsing demanding procedural steps. Finally, workflow elements such as the sequence and duration of phases and instrument usage patterns could be used to compute metrics reflective of surgical technical skills. This concept is already being applied to automate assessment during flexible endoscopy simulation, where the time taken to complete a simulated task is utilized as a metric of performance.[54] Pending correlation with clinical outcomes and other validity evidence, the same concepts of extracting workflow data could be applied to assess surgeons' performance in the OR. It is not hard to imagine a future in which such quantitative and deterministic metrics of technical skills could contribute to credentialing and privileging surgeons.

Real-time analysis of surgical workflows could greatly facilitate monitoring of intraoperative events to provide context-aware, case-specific, and timely feedback. AI models for phase recognition and instrument detection could represent the "brain" of surgical control towers, rooms from which proctors and OR managers could oversee OR activities and intervene to prevent surgical errors and inefficiencies.[55] Phase recognition and instrument detection models could continuously analyze operative videos in real-time and alert the surgical control tower when a critical step is about to be performed or when unexpected, risky deviations from normal workflows are detected. For example, a surgeon may begin to operate within a "no go" zone that raises the risk of an inadvertent injury. Proctors could respond by scrubbing into the case or providing assistance directly from the surgical control tower through telementoring and telestration. Alternatively, workflow analysis models could be used to provide direct feedback to surgeons with context-aware notifications. Such notifications could remind surgeons to implement procedure-specific best practices at the right time during interventions, potentially contributing to the overall safety of surgery. For example, notifications could remind surgeons to administer fluorescent dye, check their margins, or ensure identification of critical structures before dividing tissues. Finally, such models could compute in real-time the same workflow-based performance metrics discussed above so as to automatically provide formative feedback and coaching during procedures.

2 | LIMITATIONS AND FUTURE DIRECTIONS

Although the advances noted above are exciting and carry the promise of transforming the delivery of surgical care, it is worth noting there are several important considerations to take into account to avoid major pitfalls when applying computer vision for surgical procedures. AI is just like any other surgical innovation, and unless

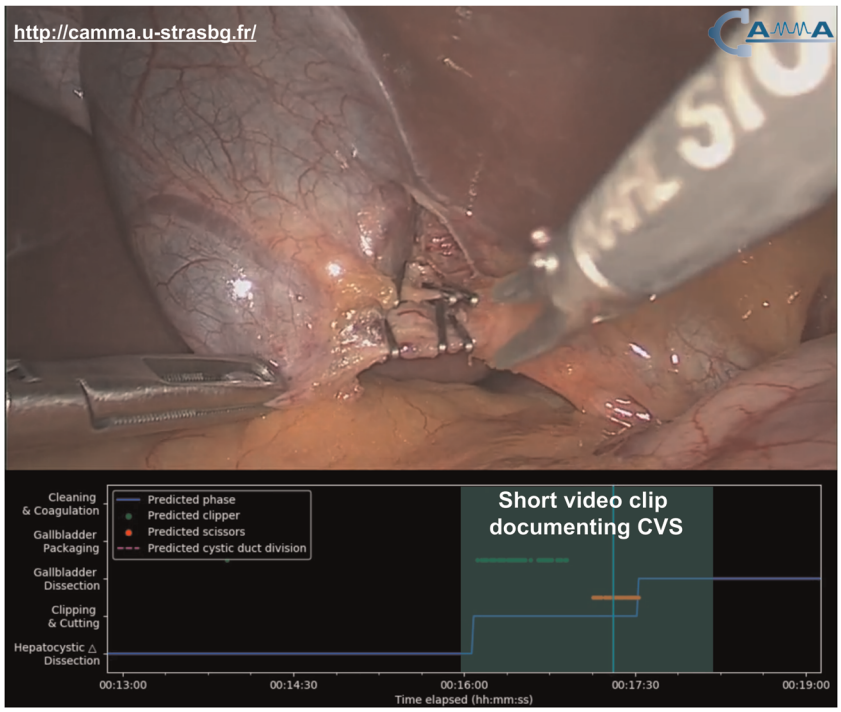

FIGURE 7 Example of EndoDigest, which uses deep learning models for workflow analysis to infer the time of the cystic duct division and extract 2:30 min long video clips for video documentation of the critical view of safety (CVS). Courtesy of CAMMA. Reproduced with permission [Color figure can be viewed at wileyonlinelibrary.com]

there is a very clear unmet need or gap in clinical practice that this technology can address, its value will be minimal and widespread adoption unlikely. DeepCVS and GoNoGoNet were specifically designed and developed to address specific cognitive behaviors which data suggests are major root causes of bile duct injuries.[23,26,27] Future models need to be similarly grounded in such data. If the aim is to develop AI models that are able to provide real-time data to surgeons based on what expert surgeons would do, it is critical to first understand how experts think and their mental processes that lead to their elite performance. Qualitative data and cognitive task analyses are a powerful method to delineate expert mental models and by gaining a better understanding of these cognitive behaviors, AI algorithms can be trained to replicate these behaviors and ultimately be deployed for real-time decision-support.[56,57]

Surgical decision-making has traditionally been presented as a linear process, where one operative step necessarily follows another and few branches in decision-making exist. For example, dissection of Calot's triangle is followed by isolation of the cystic duct or artery. More recent approaches to investigating surgical decision-making have focused on attempting to map decision-making in a manner that is more reflective of the process utilized by surgeons during the course of an operation. That is, decision making is less likely to be linear and more likely to be comprised of steps arranged in an interconnected fashion with multiple decision points that are affected by a combination of patient, surgeon, and environmental factors.[58]

An understanding of surgical decision-making plays a major role in the development of AI applications, particularly as it relates to surgical education. Decision maps that contain the possible steps of an operation can serve as the ground truth by which training data for AI is annotated. This enables applications such as automated identification of operative phases and feedback on case progression. Furthermore, an understanding of how experienced surgeons make decisions based on visuospatial data can provide clues to training algorithms to detect critical structures, operative planes, and other important visual information. Thus, surgeons can engage with researchers to better understand decision-making across a range of procedures and case complexity.

Furthermore, while it is true that models like DeepCVS or GoNoGoNet produce consistent output for any given input, they rely on human annotation to learn so their value is strictly dependent on the quality and reliability of annotations. To improve the quality of annotation, these should be performed by multiple trained reviewers using validated protocols.[59] This is likely to reduce human subjectivity and bias, especially compared to operators self-assessing during potentially stressful procedures. To further reduce operator dependencies, deep learning models should be trained on annotations from many surgeons from multiple institutions so as to learn to approximate the most commonly held interpretation for any given data point.

The notion of bias in the annotation of data also raises concerns about bias in the video data itself. Data is the life force for machine learning. With limited data, algorithms need “hints” in the form of labels to make decisions. The “cognitive boost” of labels has meant that the majority of ML/AI advances in medicine rely on supervised learning. This label reliance, though, compounds a fear of ML in general: bias. Biased or incorrect training data makes ML models (both supervised and unsupervised) output bad decisions. We currently have a limited amount of training data, so our models already have a high level of bias. In supervised learning, we compound this by having the computers learn biased human labels to categorize the already biased data. Recognizing the potential for bias is a key element in appropriately performing and interpreting ML studies, especially in medicine. Thus, while a data set may contain thousands of videos of a particular case, videos sourced from a handful of institutions or surgeons may be biased in the patient population, technique, equipment, or other factors and thus limit the generalizability of those models. Increased participation from additional centers, surgeons, and patients will be critical to ensure the success of advances in SDS.

Finally, a cultural shift will be needed in surgery and is already partially underway. A culture of quality improvement has led to an increased appetite for systematically collecting, storing, and analyzing surgical data, and this culture must now shift to include intraoperative video data as well. Surgeons must be willing to collect video data on cases they perform and should approach patients prospectively to obtain their consent to utilize such data in surgical registries. Although big data provides the means through which we are able to infer phenomena from populations, it is critical to remember that the data are collected from individuals. Patients are the beneficiaries of surgical care, and it is important to maintain the patient at the center of care whether in an operation, in an outpatient visit, or in the laboratory where SDS tools are developed. As with more traditional clinical trials conducted in oncology, early discussion with patients on the value of their surgical data is important so that patients can consider whether to contribute their operative video and other data to SDS efforts. Regulatory considerations must be taken into account and vary across countries (e.g., Health Insurance Portability and Accountability Act in the United States or General Data Protection Regulation in Europe).

3 | CONCLUSION

SDS and AI carry the potential to transform surgical training and the delivery of surgical care. Within the field of surgical education, computer vision has perhaps held the majority of the surgical attention given its easily understandable applications such as automated indexing of cases and identification of anatomy and other spatial characteristics in surgical video. Additional advances in the field will require the participation of a diverse array of surgeons, patients, and researchers to ensure that the applications of such technology are clinically meaningful.

ACKNOWLEDGMENTS

Thomas M. Ward and Daniel A. Hashimoto (DAH) receive research support from Olympus Corporation for work outside of this

manuscript. DAH is a consultant for Johnson & Johnson Institute and Verily Life Sciences and has received research support from the Intuitive Foundation for work outside of this manuscript. Nicolas Padoy is a consultant for Caresyntax and has received research support from Intuitive Surgical for work outside of this manuscript.

ORCID

Amin Madani https://orcid.org/0000-0003-0901-9851
Silvana Perretta https://orcid.org/0000-0002-5354-535X
Daniel A. Hashimoto http://orcid.org/0000-0003-4725-3104

REFERENCES

1. Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for next-generation interventions. *Nat Biomed Eng*. 2017;1:691-696.
2. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020;368:m689.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44-56.
4. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial Intelligence in Surgery: Promises and Perils. *Ann Surg*. 2018;268:70-76.
5. Navarrete-Welton AJ, Hashimoto DA. Current applications of artificial intelligence for intraoperative decision support in surgery. *Front Med*. 2020;14:369-381.
6. Goertzel B, Pennachin C. *Artificial General Intelligence*. Berlin, Heidelberg: Springer Science & Business Media; 2007.
7. Hashimoto DA. *Artificial intelligence in surgery: An AI Primer for Surgical Practice*. McGraw-Hill Education/Medical; 2020.
8. Crevier D. *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York: BasicBooks; 1993.
9. Sagiroglu S, Sinanc D. Big data: A review. *International Conference on Collaboration Technologies and Systems*. 2013;2013:42-47.
10. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444.
11. Mittal S, Vaishay S. A survey of techniques for optimizing deep learning on GPUs. *Int J High Perform Syst Archit*. 2019;99:99.
12. Russell SJ, Norvig P, Davis E. *Artificial Intelligence: a Modern Approach*. 3rd ed. Upper Saddle River: Prentice Hall; 2010.
13. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. *Nature*. 2017;550:354-359.
14. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;22:1345-1359.
15. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172:1122-31.e9
16. Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. *IEEE Trans Neural Networks Learn Syst*. 2020:1-21.
17. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*. Vol 25. New York: Curran Associates, Inc.; 2012:1097-1105.
18. Brennan TA, Leape LL, Laird NM, et al. Incidence of adverse events and negligence in hospitalized patients. *N Engl J Med*. 1991;324:370-376. https://doi.org/10.1056/nejm199102073240604
19. Baker GR. The Canadian adverse events study: the incidence of adverse events among hospital patients in Canada. *Can Med Assoc J*. 2004;170:1678-1686. https://doi.org/10.1503/cmaj.1040498
20. Forster AJ, Asmis TR, Clark HD, Al Saied G, Code CC, Caughey SC, et al. Ottawa hospital patient safety study: incidence and timing of adverse events in patients admitted to a Canadian teaching hospital. *CMAJ*. 2004;170:1235-1240.
21. Gawande AA, Thomas EJ, Zinner MJ, Brennan TA. The incidence and nature of surgical adverse events in Colorado and Utah in 1992. *Surgery*. 1999;126:66-75.
22. Madani A, Vassiliou MC, Watanabe Y, et al. What are the principles that guide behaviors in the operating room? Creating a framework to define and measure performance. *Ann Surg*. 2017;265:255-267.
23. Way LW, Stewart L, Gantert W, et al. Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. *Ann Surg*. 2003;237:460-469.
24. Madani A, Grover K, Watanabe Y. Measuring and teaching intraoperative decision-making using the visual concordance test: deliberate practice of advanced cognitive skills. *JAMA Surg*. 2019;155:78. https://doi.org/10.1001/jamasurg.2019.4415
25. Madani A, Namazi B, Altieri MS, et al. Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy [published online ahead of print November 13, 2020]. *Ann Surg*. https://doi.org/10.1097/SLA.0000000000004594
26. Brunt LM, Deziel DJ, Telem DA, et al. Safe cholecystectomy multisociety practice guideline and state of the art consensus conference on prevention of bile duct injury during cholecystectomy. *Ann Surg*. 2020;272:3-23.
27. Madani A, Watanabe Y, Feldman LS, et al. Expert intraoperative judgment and decision-making: defining the cognitive competencies for safe laparoscopic cholecystectomy. *J Am Coll Surg*. 2015;221:931-40.e8
28. Nijssen MAJ, Schreinemakers JMJ, Meyer Z, Van der Schelling GP, Crolla RMPH, Rijken AM. Complications after laparoscopic cholecystectomy: a video evaluation study of whether the critical view of safety was reached. *World J Surg*. 2015;39:1798-1803.
29. Stefanidis D, Chintalapudi N, Anderson-Montoya B, Oommen B, Tobben D, Pimentel M. How often do surgeons obtain the critical view of safety during laparoscopic cholecystectomy? *Surg Endosc*. 2017;31:142-146.
30. Mascagni P, Vardazaryan A, Alapatt D, et al. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning [published online ahead of print November 16, 2020]. *Ann Surg*. https://doi.org/10.1097/SLA.0000000000004351
31. Wauben LSGL, Van Grevenstein WMU, Goossens RHM, Meulen FHV, Lange JF. Operative notes do not reflect reality in laparoscopic cholecystectomy. *BJS*. 2011;98:1431-1436.
32. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. 2013;369:1434-1442.
33. Curtis NJ, Foster JD, Miskovic D, et al. Association of surgical skill assessment with clinical outcomes in cancer surgery. *JAMA Surg*. 2020;155:590-598.
34. Greenberg CC, Byrnes ME, Engler TA, Quamme SP, Thumma JR, Dimick JB. Association of a statewide surgical coaching program with clinical outcomes and surgeon perceptions [published online ahead of print February 10, 2021]. *Ann Surg*. https://doi.org/10.1097/SLA.0000000000004800
35. Vande Walle KA, Quamme SRP, Beasley HL, et al. development and assessment of the wisconsin surgical coaching rubric. *JAMA Surg*. 2020;155:486-492.
36. Pugh CM, Hashimoto DA, Korndorffer JR Jr. The what? How? And who? Of video based assessment. *Am J Surg*. 2021;221:13-18.
37. DaRosa DA, Zwischenberger JB, Meyerson SL, et al. A theory-based model for teaching and assessing residents in the operating room. *J Surg Educ*. 2013;70:24-30.
38. Graafland M, Schraagen JMC, Boermeester MA, Bemelman WA, Schijven MP. Training situational awareness to reduce surgical errors in the operating room. *Br J Surg*. 2015;102:16-23.

39. Padoy N. Machine and deep learning for workflow recognition during surgery. *Minim Invasive Ther Allied Technol*. 2019;28:82-90.
40. Vercauteren T, Unberath M, Padoy N, Navab N. CAI4CAI: The rise of contextual artificial intelligence in computer assisted interventions. *Proc IEEE Inst Electr Electron Eng*. 2020;108:198-214.
41. Meeuwsen FC, Van Luyn F, Blikkendaal MD, Jansen FW, Van den Dobbelsteen JJ. Surgical phase modelling in minimal invasive surgery. *Surg Endosc*. 2019;33:1426-1432.
42. Padoy N, Blum T, Ahmadi S-A, Feussner H, Berger M-O, Navab N. Statistical modeling and recognition of surgical workflow. *Med Image Anal*. 2012;16:632-641.
43. Garrow CR, Kowalewski K-F, Li L, et al. Machine learning for surgical phase recognition: a systematic review. *Ann Surg*. 2020;273:684-693. https://doi.org/10.1097/SLA.0000000000004425
44. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans Med Imaging*. 2017;36:86-97.
45. Hashimoto DA, Rosman G, Witkowski ER, et al. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. *Ann Surg*. 2019;270:414-421.
46. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Automated laparoscopic colorectal surgery workflow recognition using artificial intelligence: experimental research. *Int J Surg*. 2020;79:88-94.
47. Lalys F, Bouget D, Riffaud L, Jannin P. Automatic knowledge-based recognition of low-level tasks in ophthalmological procedures. *Int J Comput Assist Radiol Surg*. 2013;8:39-49.
48. Ward TM, Hashimoto DA, Ban Y, et al. Automated operative phase identification in peroral endoscopic myotomy [published online ahead of print July 27, 2020]. *Surg Endosc*. https://doi.org/10.1007/s00464-020-07833-9
49. Nwoye CI, Mutter D, Marescaux J, Padoy N. Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. *Int J Comput Assist Radiol Surg*. 2019;14:1059-1067.
50. Yu T, Mutter D, Marescaux J, Padoy N. Learning from a tiny dataset of manual annotations A teacher/student approach for surgical phase recognition. 2018. http://arxiv.org/abs/1812.00033. Accessed February 25, 2021.
51. Vardazaryan A, Mutter D, Marescaux J, Padoy N. Weakly-supervised learning for tool localization in laparoscopic videos. *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Cham: Springer; 2018:169-179.
52. Yamazaki Y, Kanaji S, Matsuda T, et al. Automated surgical instrument detection from laparoscopic gastrectomy video images using an open source convolutional neural network platform. *J Am Coll Surg*. 2020;230:725-32.e1
53. Mascagni P, Alapatt D, Urade T, et al. A computer vision platform to automatically locate critical events in surgical videos: documenting safety in laparoscopic cholecystectomy. *Ann Surg*. 2021. https://doi.org/10.1097/SLA.0000000000004736
54. Bencteux V, Saibro G, Shlomovitz E, et al. Automatic task recognition in a flexible endoscopy benchtop trainer with semi-supervised learning. *Int J Comput Assist Radiol Surg*. 2020;15:1585-1595.
55. Mascagni P, Padoy N. OR black box and surgical control tower: recording and streaming data and analytics to improve surgical care [published online ahead of print March 9, 2021]. *J Visc Surg*. Accepted. https://doi.org/10.1016/j.jviscsurg.2021.01.004.
56. Madani A, Watanabe Y, Vassiliou M, et al. Defining competencies for safe thyroidectomy: An international delphi consensus. *Surgery*. 2016;159(86–94):96-101.
57. Madani A, Grover K, Kuo JH, et al. Defining the competencies for laparoscopic transabdominal adrenalectomy: an investigation of intraoperative behaviors and decisions of experts. *Surgery*. 2020;167:241-249.
58. Hashimoto DA, Axelsson CG, Jones CB, et al. Surgical procedural map scoring for decision-making in laparoscopic cholecystectomy. *Am J Surg*. 2019;217:356-361.
59. Mascagni P, Fiorillo C, Urade T, et al. Formalizing video documentation of the Critical View of Safety in laparoscopic cholecystectomy: a step towards artificial intelligence assistance to improve surgical safety. *Surg Endosc*. 2020;34:2709-2714.

How to cite this article: Ward TM, Mascagni P, Madani A, Padoy N, Perretta S, Hashimoto DA. Surgical data science and artificial intelligence for surgical education. *J Surg Oncol*. 2021;124:221–230. https://doi.org/10.1002/jso.26496