Published in final edited form as:

J Am Coll Surg. 2022 June 01; 234(6): 1181–1192. doi:10.1097/XCS.0000000000000190.

Artificial Intelligence Methods and Artificial Intelligence-Enabled Metrics for Surgical Education: A Multidisciplinary Consensus

S Swaroop Vedula, MBBS, MPH, PhD,

Ahmed Ghazi, MD, FEBU, MHPE,

Justin W Collins, MBChB, MD, FRCS,

Carla Pugh, MD, PhD, FACS,

Dimitrios Stefanidis, MD, PhD, FACS, FASMBS,

Ozanan Meireles, MD,

Andrew J Hung, MD,

Steven Schwaitzberg, MD, FACS,

Jeffrey S Levy, MD, FACOG,

Ajit K Sachdeva, MD, FACS, FRCSC, FSACME, MAMSE,

Collaborative for Advanced Assessment of Robotic Surgical Skills

Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD (Vedula); the Department of Urology, University of Rochester Medical Center, Rochester, NY (Ghazi); the Division of Surgery and Interventional Science, Research Department of Targeted Intervention and Wellcome/Engineering and Physical Sciences Research Council Center for Interventional and Surgical Sciences, University College London, London, UK (Collins); the Department of Surgery, Stanford University, Stanford, CA (Pugh); the Department of Surgery, Indiana University, Indianapolis, IN (Stefanidis); the Department of Surgery, Massachusetts General Hospital, Boston, MA (Meireles); the Artificial Intelligence Center at University of Southern California Urology, Department of Urology, University of Southern California, Los Angeles, CA (Hung); UBMD Surgery, Buffalo, NY (Schwaitzberg); Institute for Surgical Excellence, Washington, DC (Levy); and Division of Education, American College of Surgeons, Chicago, IL (Sachdeva).

Abstract

Correspondence address: S Swaroop Vedula, MBBS, MPH, PhD, 3400 N Charles Street, Baltimore, MD 21218. swaroop@jhu.edu.

Drs Vedula and Ghazi contributed equally to this work.

Author Contributions

Study conception and design: Levy, Vedula, Ghazi, Collins, Martino, Feins

Acquisition of data: Vedula, Ghazi, Collins, Feins, Levy, Martino, Chen, Stefanidis, Meireles, Schwaitzberg, Hung, Stoyanov, Hernandez, Athanasiadis, Pugh, Satava, Sachdeva

Analysis and interpretation of data: Vedula, Ghazi, Collins, Hernandez, Athanasiadis, Levy

Drafting of manuscript: Vedula, Ghazi, Hernandez, Athanasiadis, Collins, Levy

Critical revision: Vedula, Ghazi, Collins, Pugh, Stefanidis, Meireles, Hung, Schwaitzberg, Levy, Sachdeva, Chen, Martino, Stoyanov, Hernandez, Athanasiadis, Feins, Satava

Members of the Collaborative for Advanced Assessment of Robotic Surgical Skills (CAARSS) who coauthored this paper are listed in the Appendix.

Support: Dr Hung is supported by the National Institutes of Health National Cancer Institute (R01CA251579–01A1).

Disclosure Information: Nothing to disclose.

BACKGROUND: Artificial intelligence (AI) methods and AI-enabled metrics hold tremendous potential to advance surgical education. Our objective was to generate consensus guidance on specific needs for AI methods and AI-enabled metrics for surgical education.

STUDY DESIGN: The study included a systematic literature search, a virtual conference, and a 3-round Delphi survey of 40 representative multidisciplinary stakeholders with domain expertise selected through purposeful sampling. The accelerated Delphi process was completed within 10 days. The survey covered overall utility, anticipated future (10-year time horizon), and applications for surgical training, assessment, and feedback. Consensus was agreement among 80% or more respondents. We coded survey questions into 11 themes and descriptively analyzed the responses.

RESULTS: The respondents included surgeons (40%), engineers (15%), affiliates of industry (27.5%), professional societies (7.5%), regulatory agencies (7.5%), and a lawyer (2.5%). The survey included 155 questions; consensus was achieved on 136 (87.7%). The panel listed 6 deliverables each for AI-enhanced learning curve analytics and surgical skill assessment. For feedback, the panel identified 10 priority deliverables spanning 2-year (n = 2), 5-year (n = 4), and 10-year (n = 4) timeframes. Within 2 years, the panel expects development of methods to recognize anatomy in images of the surgical field and to provide surgeons with performance feedback immediately after an operation. The panel also identified 5 essential that should be included in operative performance reports for surgeons.

CONCLUSIONS: The Delphi panel consensus provides a specific, bold, and forward-looking roadmap for AI methods and AI-enabled metrics for surgical education.

Surgical education is undergoing major change with training models that include objective assessments of performance and outcomes based on specific standards.[1,2] This evolution imposes new demands on training programs such as the need for multiple reliable and valid assessment tools and validated objective performance standards. The evolutionary changes in surgical education provide immense challenges and opportunities to define performance standards and to advance assessments of learners using assessment models that demonstrate strong evidence of reliability and validity. Within this context, there is tremendous potential for advancing sensor technologies that generate complex data and analytical methods. Surgical data science is a rapidly emerging domain that integrates engineering and quantitative disciplines to enable discovery and technological solutions for surgery using complex multi-modal data.[3–5] In particular, artificial intelligence (AI) methods and AI-enabled metrics allow analysis of data from multimodal sensors on cognitive, technical, physiological, and functional correlates of surgical performance and development of products that advance surgical education.[3–5] Translating products developed using AI methods and AI-enabled metrics into training curricula is nontrivial. In addition to generalizable methods, there are important considerations related to data privacy, transparency, and biases.[6] Therefore, translation of AI methods and AI-enabled metrics is facilitated best when the development is informed by needs of the stakeholders in surgical education including surgeons (educators, practitioners, and learners), professional societies and certifying boards, other regulators, and scientists in the analytical disciplines (eg engineers, statisticians). Our objective was to generate consensus among multidisciplinary stakeholders on the utility, anticipated key applications, and challenges in translating AI and metrics for surgical education.

METHOD

Systematic literature search

We systematically searched PubMed and Inspec to capture biomedical and engineering literature using index terms and keywords. The Supplemental Digital Content (http://links.lww.com/JACS/A52) shows details of the search in each database and eligibility criteria to identify relevant literature. Briefly, we included original research articles on machine learning, deep learning, and other statistical methods to compute and/or validate novel metrics or methods for surgical education (training, assessment, and feedback), in addition to studies on crowdsourcing and relevant consensus statements. We limited our search to articles published since 2016 given a past review that synthesized literature until then.[7] Two individuals independently reviewed each title and abstract from the literature search, followed by full-text review. A third author resolved disagreements in full-text review.

Didactics and moderated discussion

The workshop, held during 2 days, included presentations from 7 leading academic research groups on AI methods and AI-enabled metrics for surgical education (listed in the Supplemental Digital Content, http://links.lww.com/JACS/A52). In addition, the workshop included plenary presentations that shared technical/engineering perspectives and a presentation on a professional society's perspective on AI in surgery. Subsequently, breakout groups engaged in discussion on 1 of 3 topics for surgical education—training and education, assessment, and feedback—which was moderated by 2 panel members with surgical and engineering expertise. The breakout groups were allowed to add statements to the Delphi survey.

Delphi survey

We purposefully sampled invitations for the workshop to include domain expertise in surgery, engineering and data science, regulatory compliance, industry, and professional surgical societies, all of whom were eligible to participate in the Delphi survey. An accelerated Delphi survey was conducted in 3 rounds spanning 4 days.[6] The Delphi statements were developed based on current literature with additional statements extracted from moderated discussion among the Delphi panel, and internally piloted by a subcommittee of 5 investigators with surgical and engineering expertise. We captured participant responses using Google Forms. The survey included 5 sections: overall utility, and anticipated future during a 10-year time horizon, followed by sections on applications for surgical education and training, assessment, and feedback (Supplemental Digital Content, http://links.lww.com/JACS/A52). An 80% agreement among respondents was considered consensus.[8,9] New questions based on respondents' comments in the first round were introduced in the second round of the survey. To summarize findings and consensus, we categorized findings to address the themes shown in Table 1. Two authors independently assigned each question to a theme and resolved disagreements through discussion. We descriptively analyzed data to provide a qualitative summary of consensus among respondents.

RESULTS

Summary of literature search

A summary of our literature search to inform the survey is shown in Figure 1 (citations listed in the Supplemental Digital Content, http://links.lww.com/JACS/A52). The studies were conducted using data from simulation and the operating room, and spanned open, minimally invasive, endoscopic, and microscopic surgical techniques. The majority of included studies involved objective assessment of technical skill using different data modalities such as instrument motion, videos, neuroimaging, and so forth. Several studies focused on conventional motion metrics, both global and procedure-specific metrics, but only a few analyzed their association with patient outcomes and clinically relevant metrics in simulation.[10,11] Studies on skill assessment using machine learning mostly used deep learning methods in a black box fashion, ie without clarification about why a method yields a certain output for a given input. On the other hand, a study on feedback using machine learning incorporated domain knowledge into algorithms.[12] A few studies evaluated the effectiveness of providing metrics and other feedback derived from machine learning methods on surgical performance and learning.[13,14] Finally, several studies evaluated crowdsourcing to assess surgical skill in simulation and the operating room, and to evaluate the critical view of safety in laparoscopic cholecystectomy.[15]

Survey and characteristics of Delphi panel

The survey included a total of 155 unique questions, with 123, 46, and 24 questions in the first, second, and third rounds, respectively (Fig. 2). The Delphi panel included 40 participants. Among the 40 participants, 17 (42.5%) were surgeons, 11 (27.5%) were engineers or individuals with technical expertise in machine learning/AI, 3 (7.5%) were with the FDA, 1 (2.5%) was a lawyer, and 8 (20%) were educators and affiliates of the Society of American Gastrointestinal and Endoscopic Surgeons, American College of Surgeons, and the American College of Obstetricians and Gynecologists. Surgeons on the panel represented general surgery (47%), bariatric surgery (2.5%), urology (7.5%), gynecology (10%), and thoracic surgery (2.5%). Participants were from the US (95%) and the United Kingdom (5%). We analyzed responses from 100% of panel members in all 3 survey rounds.

Summary of panel consensus

Consensus was achieved on 136 of 155 (87.74%) questions: 102 in the first round, 23 in the second round, and 11 in the third round. Of the 19 of 155 (12.26%) questions for which consensus was not achieved, the majority were about perceived risk/benefit (6 of 19; 31.58%) and techniques for evaluating validity of AI methods and AI-enabled metrics (5 of 19; 26.32%). Agreement among the panel for questions by each theme is summarized in Table 2 (additional data in the Supplemental Digital Content, http://links.lww.com/JACS/A52).

Perceived risks/benefits with AI and metrics for surgical education

There was consensus that AI methods and AI-enabled metrics have the potential to benefit surgical education, including training (100%), assessment (97.5%), and feedback (100%).

For risks with AI and metrics, the panel only approached consensus on loss of privacy (75%) and the impact of human biases on AI (70%). Notably, there was no consensus that failure to be interpretable (57.5%) and actionable (62.5%) are potential risks with AI for surgical education.

Anticipated applications: General application areas

The eventual goal for surgical education is to enable surgeons to perform a procedure with minimal error (agreement: 97.5%) and not simply to secure operating privileges (90%). Consensus was reached that AI methods and AI-enabled metrics for surgical education should be used to assess proficiency of skill acquisition and to analyze novel data sources to define evaluation measures of surgeons' performance and to provide meaningful actionable feedback.

For high-stakes assessment of surgeons, eg initial certification or maintenance of certification, consensus on using AI methods and AI-enabled metrics was achieved only when the following criteria were met: (1) there was a high degree of evidence in favor of AI and metrics for this purpose; (2) assessments were explainable; and (3) the methods were noninferior to certified human raters. The panel agreed that AI methods and AI-enabled metrics may be used either as a screening tool before assessment by certified human raters, or to inform or supplement human assessment. For lowstakes assessment of surgeons, eg during training, the panel reached consensus that AI methods and AI-enabled metrics should supplement human assessments (97.5%), and not substitute human assessment (77.5%).

The panel agreed that AI-enhanced feedback for surgeons should include a detailed summary of their performance in the operating room and close the loop between performance in simulation settings and in the operating room by diagnosing skill deficits in the operating room, recommending personalized remediation in simulation, and monitoring learning in simulation and transfer of improved skill to the operating room.

Specific priority deliverables

Learning curves—There was unanimous agreement that AI can lead to new methodologies to analyze surgeons' learning curves, and it should be used to augment current standards for surgical proficiency. Specific deliverables on which the panel reached consensus are shown in Table 3.

Feedback for surgeons—The panel unanimously opined that AI should make it easier to deliver reliable and personalized feedback to surgical trainees and discover new metrics to ascertain when to give feedback. More broadly, the panel agreed that AI should enable active participation of the learner-surgeon in personalizing their training curricula (97.5%), for example, to help surgeons specify their learning goals. In the simulation context, AI should also enable active learning through haptic guidance in simulation.[16] Specific details of feedback that AI should provide to surgeons are shown in Table 3.

Anticipated future (10-year time horizon)

During a 10-year time horizon, there was consensus that AI methods and AI-enabled metrics can support surgical educators, for example, through detection of salient errors. The panel agreed that AI can be used to enable self-learning for surgeons, for example, through technology to guide them through personalized training curricula and to identify when feedback from an expert is needed. In addition, the panel reached consensus on several specific deliverables (Table 4).

Translating products based on AI and metrics to surgical education

Feasibility of developing products—The panel reached consensus that it was feasible to use AI methods and AI-enabled metrics to better analyze performance in simulation to predict performance in live surgery; to provide feedback on how to improve in simulation, ie close the loop between simulation and live surgery; to develop simulation models that are well mapped to live surgery; and to augment simulation with clinically relevant data.

Adoption of products—There was more than 95% agreement that AI for surgical education should be usable with interpretable output (eg provide an explanation on which aspects of input data influenced why/how the algorithm reached a certain prediction) and delivered via user-friendly interfaces. There was 80% agreement among the panel that the community should engage in appropriate branding (90%) and marketing (82.5%) to facilitate their adoption, and in protecting credit for innovations using AI outputs (87.5%). Finally, 90% of the panel agreed that AI should be used to provide feedback for surgeons on actions during the operation that may not necessarily affect patient outcomes, eg minor mishaps that may not necessarily cause adverse outcomes.

Standardization of development and application (including regulation)—The panel agreed that it is necessary to standardize how AI methods and AI-enabled metrics are used for surgical education, which currently is not standardized. In particular, 95% or more of the panel members agreed that benchmarks for performance of AI and metrics to assess surgical skill and the manner in which they are incorporated into routine practice should be standardized. Finally, the panel was unanimous that a universal annotation lexicon is a necessary step to allow adequate interoperability of AI methods and AI-enabled metrics developed using different datasets.

AI and metrics should be held accountable through regulation along the lines of FDA regulation of marketing of drugs and devices (85%), with different pathways for algorithms to be used in simulation and clinical decision support (87.5%). AI methods and AI-enabled metrics developed using simulation should be certified by an authority before they are applied in live surgery training (82.5%). Although there was 82.5% agreement that existing models can be followed to regulate AI products, our survey did not include questions on specific models. Finally, there was 95% consensus that AI should be used to inform regulation, eg by analyzing how surgeons use devices in the operating room with the intent to responsibly report events such as “near misses” or incorrect use for regulatory purposes and for designing better instruments.

Data requirements—The panel agreed that open-source, annotated datasets should be developed to facilitate and accelerate development of products using AI methods and AI-enabled metrics for surgical education. AI methods should be developed to automate annotation of datasets for subsequent use in development of methods for specific surgical applications. There should be adequate protections (including federal protections) and incentives to ease data sharing. However, the panel only approached consensus (72.5%) that datasets for AI and metrics for surgical skills assessment should be obtained through crowdsourcing.

Evaluating validity of solutions for surgical education that use AI and metrics —The panel agreed (90% or greater) that AI methods and AI-enabled metrics (1) should have strong evidence of validity to facilitate adoption by surgeons; (2) must be robust to accommodate variations in data capture, eg quality of video of the surgical field when used to assess surgical skill; and (3) should yield products that are objective, such that they are not impacted by cognitive biases that are characteristic of subjective human opinions.

The panel reached consensus that AI methods and AI-enabled metrics should be as effective (85%), and not less effective (92.5%) when compared with the current standard method for a given application. However, the panel only approached consensus that AI should be more effective than the standard method (77.5%). In fact, subsequent questions to explore this opinion revealed no consensus on this question for the simulation (62.5%) and operating room settings (70%).

The panel reached consensus on nearly all items on the ground truth against which to benchmark AI and metrics for skill assessment, feedback, learning curves, and personalized curricula (Table 5). To validate AI and metrics for surgical skill assessment, the panel reached consensus that the ground truth benchmarks for both skill assessment and feedback should be provided by expert surgeons (90%). However, benchmarking against expert assessments is susceptible to biases inherent in subjective human opinions (85%). On the other hand, effectiveness of feedback using AI and metrics should be evaluated against expert feedback on learners' performance in the operating room (100%).

Finally, the panel addressed the role of simulation in validating AI methods and AI-enabled metrics for surgical education. The panel agreed (87.5%) that methods for application in live surgery should be tested in a controlled simulation setting that allows standardization of confounders. Specifically, they reached consensus (95%) that new products using AI methods and AI-enabled metrics to inform surgical decisions must be evaluated in high-fidelity simulations before their use in the operating room, but only approached consensus (77.5%) on products using AI and metrics to predict patient outcomes.

DISCUSSION

Our findings provide a specific roadmap for how AI methods and AI-enabled metrics can be used to advance surgical education. Surgical data science offers innovative methods, including AI methods and AI-enabled metrics, to analyze surgical performance. However, advances in technology for surgical education have been constrained by lack of consensus on clinical priorities. Our survey identified specific applications and priority deliverables, which reflect consensus among a broad range of stakeholders, including surgeons, educators, engineers, and experts affiliated with professional surgical societies and regulatory agencies. Therefore, our findings capture the surgical education community's expectations of how AI methods and AI-enabled metrics can be used to translate research in surgical data science into products that benefit surgical education.

The ultimate goal of surgical education is to develop surgeons' psychomotor and cognitive skills to effectively perform operations with the best possible outcomes for patients. Training in the operating room is irreplaceable, but it should be provided while optimizing patient care and safety. This may be achieved in many ways, pre-, intra-, and postoperatively, including maximizing skills that can be acquired through simulation before the operating room, objective screening assessments for participation in the operating room, intraoperative coaching and guidance using real-time performance data, early warning systems to detect impending errors, objective postprocedure assessments, and feedback to enable deliberate practice. For the educators, professional societies, and certifying bodies, objective measures of trainee surgeons' skill may be available in an accessible and interpretable form (eg a surgical portfolio) that allows reproducible comparative analytics across surgeons. Although the preceding narrative describes a transformative vision of surgical education compared with traditional training models,[17] the consensus achieved by the Delphi panel indicates its current relevance and feasibility through AI methods and AI-enabled metrics.

Despite perceived feasibility of surgical education enhanced by AI methods and AI-enabled metrics, many technical solutions that enable priority deliverables (Table 4) must be developed for it to become reality. These solutions address a wide range of problems including assessment, surgical scene analysis (object recognition, semantic segmentation, ie delineating semantic structures in data), and process or workflow analysis. Although the current literature shows an emphasis on methods for skill assessment, surgical scene analysis is a foundational technology that enables several priority deliverables.[18,19] For example, object recognition and segmentation of objects in video images are necessary to discover patterns in data that are useful to provide feedback to surgeons. Despite much research on recognizing objects in surgical video images,[20–26] it may not yet be considered a solved problem because of limited proof of generalizability.[26] Whereas past research on semantic segmentation in surgical videos was limited to instruments, recent studies reported methods to segment anatomy and instruments.[18,25,27,28] Similarly, recognition of steps in surgical procedures is a widely tackled problem using videos. Although most studies report algorithms developed and validated with the same dataset, ie internal validation,[23–25,27,28] evidence on external validity of the algorithms is lacking. Generalizable methods for detailed analysis of data on the surgical process are necessary to achieve some of the priority deliverables identified by the panel in the near term.

Although assessment of surgeons is a potential utility of AI methods, current evidence shows that the technology is not ready for routine use. The panel consensus cited a high degree of evidence, explainability, and noninferiority to certified human raters as criteria for AI to be useful for high-stakes assessment of surgeons. Currently, there is limited evidence of external validity of AI methods to assess skills using instrument motion, video, and eye tracking data.[7,10,11,29–33] Although research in simulation settings has explored explainability of assessments from AI methods,[33] they are not well developed, and similar findings using operating room data are lacking. Finally, validity of AI methods must be demonstrated using unbiased datasets that are representative and that support adequately powered analyses.

Some priority deliverables identified by the panel require interaction between AI technologies and the surgeon, eg intraoperative navigation, and guidance on next steps after an error or on optimal use of instruments. These high-value deliverables require advances in multiple areas of data analysis that complement each other and that are possible through mechanisms for sustained research funding. Applications that require interaction between technology and the surgeon should be evaluated for safety and effectiveness before they are adopted in routine patient care. Although randomized controlled trials are useful for evaluating AI-enhanced technologies,[14] simulation can also play an important role in evaluation and translation of AI technologies for surgical education. The role of simulation for this purpose should be clarified in future research. Lack of consensus among the panel on some items clarifies current expectations from AI methods and AI-enabled metrics within the surgical education community. For instance, the panel did not reach consensus on whether AI and metrics should surpass the accuracy of human raters for skill assessment. This is unlike clinicians' expectations for other applications such as AI-assisted radiologic diagnosis.[34] We also anticipate that expectations of the surgical education community may evolve with emerging evidence. For example, the panel did not reach consensus that grading procedure difficulty is a priority deliverable for AI methods and AI-enabled metrics. In fact, there is minimal research on accuracy of AI to predict procedure difficulty grading, although a recent study evaluated algorithms to detect the critical view of safety in cholecystectomy procedures in patients with different grades of disease severity.[35] Evidence of the association between disease severity and procedure difficulty may inform revised priorities for AI methods and AI-enabled metrics in surgical education.[36] Although our findings present a unique insight and a specific roadmap for advances in AI methods and AI-enabled metrics for surgical education, our study has limitations. For example, we did not explore all items on which the panel did not reach consensus. Specifically, the panel only approached consensus on crowdsourcing for assessment of proficiency in skill acquisition. We did not follow up with questions on different settings in which these assessments are performed and used, eg low stakes vs high stakes, simulation vs operating room, and basic skills vs full procedures. However, recent literature provides some insights into the surgical community's concerns about crowdsourcing, especially for high-stakes assessments.[36] We did not explore how the surgical community can lead the innovation. However, a past consensus statement from a multinational group of stakeholders, with a greater representation from engineers than surgeons, opined that surgical data science must be developed as a career path for both independent scientists and surgeon-scientists.[3,4] In fact, none of the participants practicing surgery had advanced training in computer science. Integrating surgical data science into the medical student or surgical training curricula should be further explored with relevant stakeholders.

Our survey makes it abundantly clear how AI methods and AI-enabled metrics can lead to innovations that spur progress in surgical education. There is increasing societal commitment to the role of technology in education throughout surgeons' careers. For example, the surgical community is actively exploring ways to integrate video-based assessments of surgeons during training, and for purposes of initial certification and maintenance of certification.[37] There is a vibrant engineering research community willing to engage with surgeons and advance surgical data science.[3,4] The missing ingredients are data, annotations, funding streams, and translation of research. Free access to well-annotated data are critical to achieve the deliverables identified in this study through AI methods and AI-enabled metrics. In fact, access to annotated data may be the most important determinant of whether and to what extent AI methods and AI-enabled metrics can have a transformative impact on surgical education. Efforts by the Society of American Gastrointestinal and Endoscopic Surgeons AI Task Force are addressing the data and annotation challenges through consensus building and multicenter research initiatives.[38] However, these initiatives represent major but early steps that should be supplemented by efforts by individual researchers and collaborative consortia.

CONCLUSIONS

As data on surgical performance become ubiquitously available through sensing technologies and simulation, AI methods and AI-enabled metrics are positioned to play a pivotal role in the future of surgical education. Consensus among the Delphi panel in this study lays out a bold and forward-looking roadmap of expectations of how AI methods and AI-enabled metrics can drive progress in surgical-education with specific deliverables that include measuring learning curves, assessment of skill, and technology to provide surgeons with feedback.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment:

We thank participants in the Delphi survey for their opinions.

Disclosures outside the scope of this work: Dr Ghazi is a paid consultant to Olympus, Medtronic, and Johnson & Johnson; and receives research funding from Intuitive Surgical, Inc, Coloplast, and Lumines. Dr Collins is employed by CMR Surgical. Dr Meireles is a paid consultant to Olympus, receives research funding from Olympus, and receives grant funding from CRICO. Dr Stoyanov is employed by Digital Surgery, Ltd, and holds stock/stock options in Odin Vision, Ltd. Dr Martino is a patient safety consultant for Medtronic, Intuitive Surgical, Inc, and Johnson & Johnson, has been paid for his expert testimony, and receives speaker's fees from Glaxo Smith Kline. Dr Stefanidis's institute is supported by research funding from Becton Dickinson and Intuitive Surgical, Inc. Other authors have nothing to disclose.

APPENDIX

Members of the Collaborative for Advanced Assessment of Robotic Surgical Skills (CAARSS): Danail Stoyanov, PhD, Department of Computer Science, and Division of Surgery and Interventional Science, Research Department of Targeted Intervention, University College London, London, UK; Chi Chiung Grace Chen, MD, MHS, FACOG, Department of Gynecology and Obstetrics, Johns Hopkins University School of Medicine, Baltimore, MD, and Malone Center for Engineering in Healthcare, Johns Hopkins

University, Baltimore, MD; Edward Hernandez, MD, Department of Surgery, Indiana University, Indianapolis, IN; Dimitrios I Athanasiadis, MD, Department of Surgery, Indiana University, Indianapolis, IN; Martin A. Martino, MD, FACS, FACOG, Division of Gynecologic Oncology, Ascension Health, Jacksonville, FL; Richard Feins, MD, University of North Carolina, Chapel Hill, NC; Richard Satava, MD, FACS, Department of Surgery, University of Washington, Seattle, WA.

REFERENCES

1. Sonnadara RR, Mui C, McQueen S, et al. Reflections on competency-based education and training for surgical residents. J Surg Educ 2014;71:151–158. [PubMed: 24411437]
2. Carraccio C, Wolfsthal SD, Englander R, et al. Shifting ods and AI-enabled metrics can drive progress in surgical paradigms: from Flexner to competencies. Acad Med 2002;77:361–367. [PubMed: 12010689]
3. Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for next-generation interventions. Nat Biomed Eng 2017;1:691–696. [PubMed: 31015666]
4. Maier-Hein L, Eisenmann M, Sarikaya D, et al. Surgical data science from concepts toward clinical translation. Med Image Anal 2022;76:102306.
5. Vedula SS, Hager GD. Surgical data science: the new knowledge domain. Innov Surg Sci 2017;2:109–121. [PubMed: 28936475]
6. Collins JW, Marcus HJ, Ghazi A, et al. Ethical implications of AI in robotic surgical training: a Delphi consensus statement. Eur Urol Focus 2021 Apr 30:S2405–4569(21)00112–7.
7. Vedula SS, Ishii M, Hager GD. Objective assessment of surgical technical skill and competency in the operating room. Annu Rev Biomed Eng 2017;19:301–325. [PubMed: 28375649]
8. Cakir OO, Castiglione F, Tandogdu Z, et al. Management of penile cancer patients during the COVID-19 pandemic: an eUROGEN accelerated Delphi consensus study. Urol Oncol 2021;39:197.e9–197.e17.
9. Collins JW, Ghazi A, Stoyanov D, et al. Utilising an accelerated Delphi process to develop guidance and protocols for telepresence applications in remote robotic surgery training. Eur Urol Open Sci 2020;22:23–33. [PubMed: 34337475]
10. Hung AJ, Chen J, Gill IS. Automated performance metrics and machine learning algorithms to measure surgeon performance and anticipate clinical outcomes in robotic surgery. JAMA Surg 2018;153:770–771. [PubMed: 29926095]
11. Witthaus MW, Farooq S, Melnyk R, et al. Incorporation and validation of clinically relevant performance metrics of simulation (CRPMS) into a novel full-immersion simulation platform for nerve-sparing robot-assisted radical prostatectomy (NS-RARP) utilizing three-dimensional printing and hydrogel casting technology. BJU Int 2020;125:322–332. [PubMed: 31677325]
12. Holden MS, Xia S, Lia H, Keri Z, et al. Machine learning methods for automated technical skills assessment with instructional feedback in ultrasound-guided interventions. Int J Comput Assist Radiol Surg 2019;14:1993–2003. [PubMed: 31006107]
13. Yang YY, Shulruf B. Expert-led and artificial intelligence (AI) system-assisted tutoring course increase confidence of Chinese medical interns on suturing and ligature skills: prospective pilot study. J Educ Eval Health Prof 2019;16:7. [PubMed: 30986892]
14. Buescher JF, Mehdorn AS, Neumann PA, et al. Effect of continuous motion parameter feedback on laparoscopic simulation training: a prospective randomized controlled trial on skill acquisition and retention. J Surg Educ 2018;75:516–526. [PubMed: 28864265]
15. Deal SB, Stefanidis D, Telem D, et al. Evaluation of crowdsourced assessment of the critical view of safety in laparoscopic cholecystectomy. Surg Endosc 2017;31:5094–5100. [PubMed: 28444497]
16. Marchal Crespo L, Reinkensmeyer DJ. Haptic guidance can enhance motor learning of a steering task. J Mot Behav 2008;40:545–556. [PubMed: 18980907]
17. Halstead WS, Stanton A. Friedberg MD. Surgical papers. Rare Book Collection of Rush University Medical Center at the University of Chicago. 1924.

18. Grammatikopoulou M, Flouty E, Kadkhodamohammadi A, et al. CaDIS: Cataract dataset for surgical RGB-image segmentation. Med Image Anal 2021;71:102053.

19. Neumuth T, Meissner C. Online recognition of surgical instruments by information fusion. Int J Comput Assist Radiol Surg 2012;7:297–304. [PubMed: 22005841]

20. 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;230:725–732.e1. [PubMed: 32156655]

21. Zisimopoulos O, Flouty E, Luengo I, et al. DeepPhase: Surgical Phase Recognition in CATARACTS Videos. In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G (eds). Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science, vol 11073. Springer, Cham.

22. Qin F, Li Y, Su Y-H, Xu D, Hannaford B. Surgical Instrument Segmentation for Endoscopic Vision with Data Fusion of CNN Prediction and Kinematic Pose. 2019 International Conference on Robotics and Automation (ICRA); 2019: 9821–9827. Available at: 10.1109/icra.2019.8794122.

23. Ward TM, Mascagni P, Ban Y, et al. Computer vision in surgery. Surgery 2021;169:1253–1256. [PubMed: 33272610]

24. Ward TM, Mascagni P, Madani A, et al. Surgical data science and artificial intelligence for surgical education. J Surg Oncol 2021;124:221–230. [PubMed: 34245578]

25. Bilgic E, Gorgy A, Yang A, et al. Exploring the roles of artificial intelligence in surgical education: a scoping review. Am J Surg 2021 Nov 30:S0002–9610(21)00682–6.

26. Sokolova N, Schoeffmann K, Taschwer M, et al. Evaluating the Generalization Performance of Instrument Classification in Cataract Surgery Videos. In: Ro YM, Cheng W-H, Kim J, et al. (eds). MultiMedia Modeling. MMM 2020. Lecture Notes in Computer Science, vol 11962. Springer, Cham.

27. Garrow CR, Kowalewski KF, Li L, et al. Machine learning for surgical phase recognition: a systematic review. Ann Surg 2021;273:684–693. [PubMed: 33201088]

28. Anteby R, Horesh N, Soffer S, et al. Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. Surg Endosc 2021;35:1521–1533. [PubMed: 33398560]

29. Kim TS, O'Brien M, Zafar S, et al. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. Int J Comput Assist Radiol Surg 2019;14:1097–1105. [PubMed: 30977091]

30. Ahmidi N, Poddar P, Jones JD, et al. Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. Int J Comput Assist Radiol Surg 2015;10:981–991. [PubMed: 25895080]

31. Ahmidi N, Hager GD, Ishii L, et al. Surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery: surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery. Med Image Comput Comput Assist Interv 2010;13(pt 3):295–302. [PubMed: 20879412]

32. Ahmidi N, Ishii M, Fichtinger G, et al. An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data. Int Forum Allergy Rhinol 2012;2:507–515. [PubMed: 22696449]

33. Ismail Fawaz H, Forestier G, Weber J, et al. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. Int J Comput Assist Radiol Surg 2019;14:1611–1617. [PubMed: 31363983]

34. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15:e1002686.

35. Korndorffer JR Jr, Hawn MT, Spain DA, et al. Situating artificial intelligence in surgery: a focus on disease severity. Ann Surg 2020;272:523–528. [PubMed: 33759839]

36. Williams RG, George BC, Bohnen JD, et al. A proposed blueprint for operative performance training, assessment, and certification. Ann Surg 2021;273:701–708. [PubMed: 33201114]

37. Pugh CM, Hashimoto DA, Korndorffer JR Jr. The what? how? and who? of video based assessment. Am J Surg 2021;221:13–18. [PubMed: 32665080]

38. Meireles OR, Rosman G, Altieri MS, et al. ; SAGES Video Annotation for AI Working Groups. SAGES consensus recommendations on an annotation framework for surgical video. Surg Endosc 2021;35:4918–4929. [PubMed: 34231065]

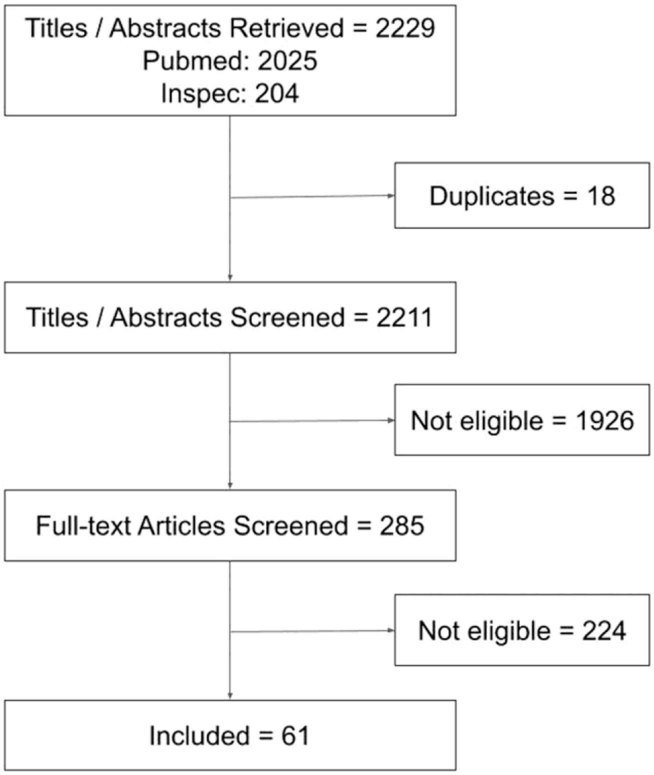
Titles / Abstracts Retrieved = 2229
Pubmed: 2025
Inspec: 204
Duplicates = 18
Titles / Abstracts Screened = 2211
Not eligible = 1926
Full-text Articles Screened = 285
Not eligible = 224
Included = 61

Figure 1.

Results of literature search.

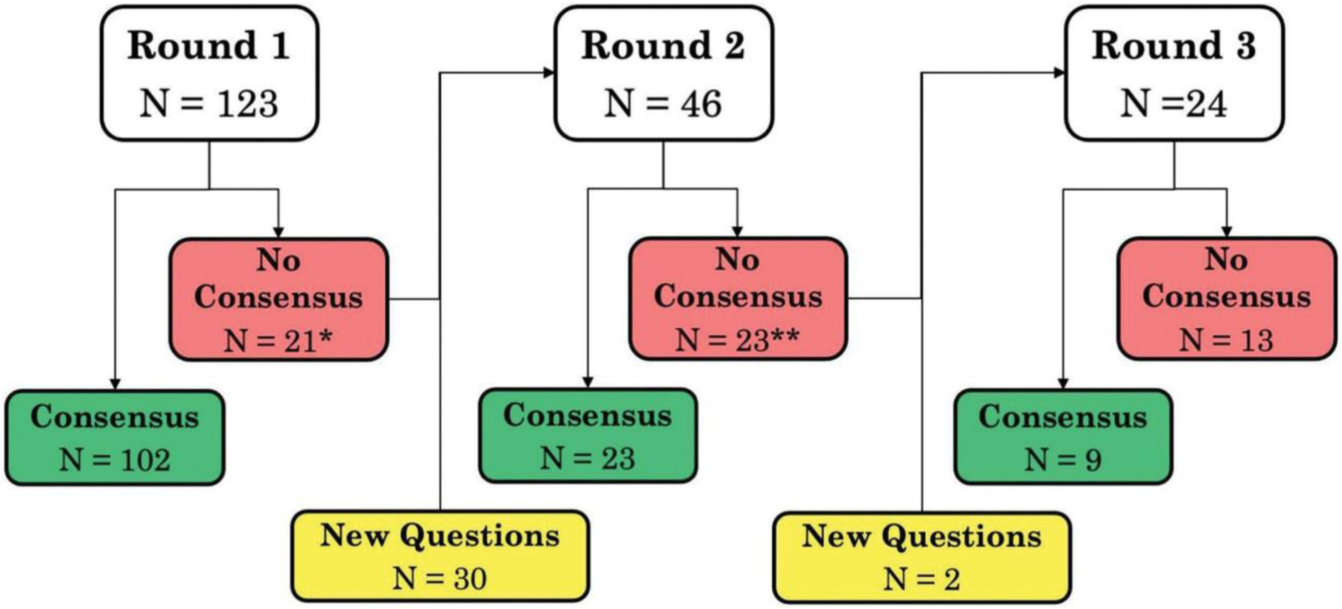

Figure 2.

Flow of questions across 3 rounds of the survey. *Count includes 5 questions with no consensus that were omitted and replaced with new questions. **Count includes 1 question with no consensus that was omitted and replaced with a new question.

Table 1

Theme Questions Used to Summarize Consensus from the Delphi Panel

Theme question
1. What are the perceived risks/benefits with AI and metric s for surgical education?
2. What are the anticipated applications/future for AI and metrics for surgical education?
a. General application areas
b. Specific priority deliverables
c. Anticipated future (10-year time horizon)
3. What will it take to translate products based on AI and metrics into surgical education?
a. Feasibility of developing products
b. Adoption of products
c. Standardization ofdevelopment pathway (including regulation)
d. Data requirements to develop products (using simulation, using crowdsourcing, and data sharing)
e. Evaluating validity ofproducts for surgical education that use AI and metrics

AI, artificial intelligence.

Table 2

Agreement Among the Panel for Questions in Each Theme in Our Survey

		Majority response to specific question with consensus				Majority response to specific question without consensus			
		Yes		No		Yes		No	
Theme	Questions in survey, n	No. of questions	Median agreement, % (IQR)	No. of questions	Median agreement, % (IQR)	No. of questions	Median agreement, % (IQR)	No. of questions	Median agreement, % (IQR)
Q1. Risk/benefit	14	7	97.5 (97.5–100)	1	90*	5	70 (62.5–75)	1	65*
Q2a. General application areas	15	13	92.5 (87.5–95)	1	80*	0	NA	1	77.5*
Q2b. Specific priority deliverables	46	41	95 (87.5–97.5)	0	NA	3	77.5 (72.5–77.5)	0	NA
Q2c. Anticipated future (10-year horizon)	16	14	86.25 (85–89.3)	0	NA	2	65, 70*	0	NA
Q3a. Feasibility of developing products	8	8	85 (85–87.5)	0	NA	0	NA	0	NA
Q3b. Adoption of products	9	9	90 (87.5–97.5)	0	NA	0	NA	0	NA
Q3c. Standardization of development pathway	12	10	92.5 (85.6–96.9)	1	90*	1	75*	0	NA
Q3d(i). Data requirements to develop products— data from simulation	1	1	100	0	NA	0	NA	0	NA
Q3d(ii). Data requirements to develop products— data from crowdsourcing	1	0	NA	0	NA	1	72.5	0	NA
Q3d(iii). Data requirements to develop products—data sharing	3	3	97.5 (97.5–97.5)	0	NA	0	NA	0	NA
Q3e. Evaluating validity	32	26	90 (85.6–95)	1	92.5*	5	77.5 (70–77.5)	0	NA

* Median and interquartile range (IQR) are reported when more than 2 questions are in the category. When there is a single question in the category, the agreement (%) on that question is reported. When 2 questions are in the category, the agreement on each question (%) is shown, respectively.

NA, not applicable.

Table 3

Deliverables for Artificial Intelligence and Metrics for Surgical Education with Panel Consensus

Deliverable; metric
Learning curve (panel agreement 80% or more)
1. Predict where along the learning curve individual surgeons fall at any time within a curriculum, including at the start and end of learning
2. Predict lack ofprogress along the learning curve (i.e., arrested learning)
3. Recommend effective alterations in curriculum to pre-empt failure to progress and to remediate arrested learning
4. Personalize predictions through analysis ofpatterns across large numbers of surgeons
5. Standardize measures of learning curves applicable to categories of similar procedures (e.g., reconstructive or ablative procedures)
6. Identify metrics to include in a surgeon's portfolio, which captures longitudinal measures ofprespecified metrics.
a. Metrics that have greatest impact on surgical performance because they are associated with a clinical outcome or training in simulation to improve on them is associated with better clinical outcomes
b. Metrics that accurately predict the end point (i.e., competency level) ofthe learning curve
Surgical skill assessment (panel agreement 90% or more)
1. Assess skill at end-of-rotation and end-of-year within training curricula
2. Assess each procedure surgeons perform during training
3. Assess technical skill in simulation and operating room in real-time (as data are captured) and offline (analyze recorded data after operation is performed)
4. Assess technical skill at more granularity than for a procedure (e.g., steps)
5. Deconstruct surgical activities to facilitate granular skill assessment
6. Assess nontechnical skills in the operating room offline but not in real time (77.5% agreement)
Feedback for surgeons in the operating room (panel agreement 87.5% or more)
1. Provide summary report on operation with the following information:
a. Avoidance of "near misses" based on artificial intelligence to identify anatomical structures, e.g., prevent unintended injury to underlying structures
b. Detection of error (minor and major)
c. Illustration of possible actions for the surgeon to recover from an error
d. Assessment of identification of critical portions of the operation, e.g., identify critical view of safety in laparoscopic cholecystectomy
e. F eedback on how surgeons can best use devices/instruments
2. Formative feedback in real-time

Table 4

Specific Priority Deliverables Related to Feedback That Artificial Intelligence Methods and Artificial Intelligence—Enabled Metrics for Surgical Education Should Meet During the Next 10 Years

Priority deliverable
Short-term (2-year time frame)
1. Recognize anatomy in images from videos of the surgical field (97.5%)
2. Provide performance feedback to surgeon immediately after the operation (85%)
Mid-term (5-year time frame)
3. Identify parts of the operation on which the surgeon needs feedback (82.5%)
4. Overlay images to display surrounding anatomy (90%)
5. Guide surgeons on expo sure of the surgical field (e.g., artificial intelligence-guided cardiac ultrasound for noncardiologists) (82.5%)
6. Guide surgeons on optimal use of instruments/devices (85%)
Long-term (10-year time frame)
7. Enable intraoperative navigability using video, kinematics, and other imaging data for multiple procedures (eg navigation in sinus surgery using CT imaging) (85%)
8. Detect intraoperative error (82.5%)
9. Provide guidance on the next best step to address an intraoperative error or complication (87.5%)
10 Grade difficulty of surgical procedure (65%; no consensus)

Panel agreement shown in parentheses.

Table 5

Variables That Can Be Used as Ground Truth to Validate Artificial Intelligence Methods and Artificial Intelligence— Enabled Metrics for Learning Curves and Personalized Curricula

Variable
Skill assessment
1. Skill category (eg expert/novice/intermediate; 85%)
2. Standardized structured rating scales (95%)
3. Patient outcomes (90%)
Learning curve
1. Specific operative process measures, such as blood loss, ischemia time, and so forth (90%), but not operative time (77.5%)
2. Measures of procedure-specific surgical success, such as continence or nonconversion (87.5%)
3. Postoperative outcomes such as complication; length ofhospital stay (92.5%)
4. Oncologic outcomes such as surgical margins, number of lymph nodes, and so forth (95%)
5. Patient-specific outcomes such as survival, patient-reported outcomes such as quality of life, satisfaction, and so forth (82.5%)
Personalized curricula
1. Standardized milestones such as achieving a certain level of skill (100%)
2. Surgeon perception of whether learner can be entrusted with specific aspects of care (80%)
3. Performance in the operating room (100%)
4. Surgical outcomes in patients (95%)
5. Error in performing the operation (100%)

Panel agreement shown in parentheses.